University of Szeged

Doctoral School in Linguistics

PhD Program in Theoretical Linguistics

Danang Satria Nugraha

A Theoretical and Corpus Linguistics Study of the Light Verb Constructions: Empirical Data from Indonesian

PhD Dissertation

Supervisor:
Veronika Vincze, PhD

Szeged, 2025

DISSERTATION DECLARATION

I solemnly declare that the research presented herein is the product of my original intellectual endeavor, conducted under the tutelage of Veronika Vincze, PhD. I attest that no segment of this dissertation has been submitted for academic consideration towards any degree or qualification at any institution. All external materials are meticulously attributed and referenced. Save for the integration of content from my concurrent publications, this dissertation is entirely my independent work. I consent to the public availability of the final version of this thesis through the university's research repository, institutional channels, and relevant search engines.

Danang Satria Nugraha
Szeged, April 23, 2025

Table of contents

Chapter 1 Introduction

Chapter 2 Method of the study

Chapter 3 Frequency and distribution of Indonesian LVCs

Chapter 4 Verb element within Indonesian LVCs

Chapter 5 Noun element within Indonesian LVCs

Chapter 6 Conclusion

List of tables

List of figures

List of Abbreviations

1	first person
2	second person
3	third person
A	agent-like argument of canonical transitive verb
AB	abstract
AF	affixed form
ACC	accusative
ADJ	adjective
ADV	adverb(ial)
ADP	ad position
An	animal (prototypical meaning of word)
Ac	action (prototypical meaning of word)
Ba	base form
BEN	benefactive
BN	basic physiological needs
BL	blow
Co	concrete from
CO	cognitive process
COMPL	complement
CP	counterpart
ES	emotional states
FI	fixed valency
FE	flexible valency
GE	gesture
II	inherently intransitive
INF	infinitive
IDC	The Indonesian Web Corpus (idTenTen)
ILCC	Indonesian – Leipzig Corpora Collection
IWC	The Indonesian web corpus (idWaC)
IT	inherently transitive
KS	Kolmogorov–Smirnov
KDE	kernel density estimate
LVC	light verb construction
LRL	low resource language
Ma	material (prototypical meaning of word)
MS	mental states
MO	motion
N	noun
NB	Naïve Bayes classifier
NC	non-verbal communication

NCP	no counterpart
OBJ	object
Ob	objects (prototypical meaning of word)
P	patient-like argument of canonical transitive verb
PA	physical activities
PASS	passive
PE	people (prototypical meaning of word)
PC	physical sensations
PCA	principal component analysis
PI	physical states
PL	plural
Pl	plants (prototypical meaning of word)
PMW	per million words
POSS	possessive
PRED	predicative
PROG	progressive
PROPN	proper name
PRS	present
Pr	processes (prototypical meaning of word)
PST	past
PURP	purposive
Qu	qualities (prototypical meaning of word)
REFL	reflexive
RES	resultative
SD	standard deviation
SE	sensations
SBJ	subject
SG	singular
SLIC	SEAlang Library Indonesian Text Corpus
SVM	support vector machine
TOP	topic
TLV	true light verb
TR	transitive
V	verb
VAV	vague action verb
VC	verbal communication
VS	vocal sounds
Σ	total

Preface

My initial foray into the analysis of verb constructions began in 2014 with my undergraduate thesis, where I explored the role of verbs in marking theme-rheme structure. This interest further developed in my 2016 master's thesis, which focused on the analysis of denominal verb constructions. This enduring fascination with verb constructions ultimately led to the present doctoral dissertation, an investigation into the properties of light verb constructions in Indonesian.

This dissertation examines light verb constructions in Indonesian, my native language, spoken by nearly 275 million people worldwide. While I acknowledge the possibility of prior research on this topic, to the best of my knowledge, dedicated studies on light verb constructions in Indonesian remain scarce. Despite this academic lacuna, such constructions are prevalent in everyday Indonesian, with examples like '*memberikan kuliah*' (to give a lecture), '*mengambil kesempatan*' (take a chance), and '*membuat keputusan*' (make a decision) as well as other light verb constructions readily attesting to their common usage. This research aims to address this gap in the literature by providing a detailed account of this pervasive linguistic phenomenon.

This research offers an explanatory description of Indonesian light verb constructions, drawing upon theoretical linguistics and corpus-based methodology. Key chapters include analyses of the frequency and distribution of these constructions, along with a detailed examination of their constituent elements. The verbal element, as the initial component, is scrutinized to provide a comprehensive understanding of the phenomenon. Complementing this, a thorough analysis of the nominal element is undertaken, recognizing its crucial role in elucidating the complete picture. These perspectives form the foundation for the comprehensive analysis presented herein.

While my intuitions as a native speaker of Indonesian provide a valuable starting point, they are insufficient for a rigorous investigation. To overcome this limitation, I employ corpus linguistics methods, which enable both quantitative and qualitative analysis of large-scale data. Furthermore, by leveraging machine learning algorithms, I provide quantitative evidence for the intricate semantic relationship between verbs and nouns within Indonesian light verb constructions. This approach sheds light on the dynamic interplay of semantic forces that characterize these constructions.

In sum, this research, while acknowledging its limitations, aspires to embody the wellspring articulated by Otto Jespersen, "Any linguistic phenomenon may be regarded either from without or from within, either from the outward form or from the inner meaning" (1924: 33). To some extent, by combining introspective linguistic knowledge with the corpus-based analysis this dissertation provides a detailed exploration of Indonesian light verb constructions.

Acknowledgement

This dissertation would not have been possible without the invaluable assistance and support I received throughout my doctoral studies. My academic journey began in mid-2021, when I was fortunate enough to become an advisee of Veronika Vincze, PhD within the Doctoral School of Linguistics (DSL) at the University of Szeged (SZTE). I am deeply indebted to her for her guidance and mentorship, which profoundly shaped my academic development during these past four formative-years.

This opportunity was facilitated by Professor Dr. Enikő Németh T., DSc., Head of the DSL, whose role in my admission to this esteemed program was paramount. I extend my sincere gratitude to her for recognizing my potential and granting me the chance to pursue my doctoral studies at SZTE. I also wish to acknowledge the contributions of the entire DSL faculty. Associate Professor Dr. Lívia Ivaskó, PhD, Programme Director of Theoretical Linguistics, provided consistent and invaluable feedback on my progress throughout my studies. Tibor Szécsényi, PhD generously welcomed me into his workspace and included me in several funded conferences. Dr. Małgorzata Suszczyńska's insightful lectures and discussions inspired me to develop an assignment into a publishable paper, which was published as a chapter by a renowned international publisher. I am also grateful for the stimulating conversations with Dr. Gyula Zsigri, both inside and outside the classroom, on the contemporary challenges faced by language researchers. Martina Katalin Szabó, PhD's computational linguistics course, with its access to Python in the virtual lab, proved to be an invaluable resource. Finally, I thank Zsuzsanna Németh, PhD and Csenge Guba for sharing their knowledge on the history of linguistic studies in Szeged and Hungary.

I would be remiss not to acknowledge the significant contribution of Professor Károly Bibok, who fortified my theoretical foundation for this research. His advice during my comprehensive examination provided not only a solid base but also the cornerstone upon which this dissertation was built. I also extend my gratitude to Enikő Héja, PhD for her guidance on the corpus-based methodology employed in this study.

My doctoral studies, including the research culminating in this dissertation, were generously supported by the Ministry of Foreign Affairs and Trade of the Hungarian Government through the prestigious Stipendum Hungaricum Scholarship, managed by the Tempus Public Foundation. This endeavor also benefited from the support of the Ministry of Higher Education, Research, and Technology of the Republic of Indonesia. Furthermore, I acknowledge the fundamental support provided by Sanata Dharma University, Indonesia, my home institution to which I will return upon completion of my studies.

I am deeply grateful for the unwavering support of my colleagues and family. To my colleagues within and outside the Department of Indonesian Language & Education at Sanata Dharma University, I express my sincere appreciation for their encouragement and camaraderie. Last but certainly not least, to my parents, my spouse, my siblings, and my nephew, thank you for your endless love and support, which sustained me throughout this journey.

Szeged, April 23, 2025

Abstract

This dissertation investigates Indonesian Light Verb Constructions (LVCs) through the integrated lenses of theoretical linguistics and corpus-based analysis. The primary part of the study provides an empirical foundation, drawing on four corpora to assess the frequency and distribution of LVCs. K-means clustering reveals three natural groupings (low-, medium-, and high-frequency), with Hypothesis 1 supported by significant cross-corpus rank consistency (Spearman's $r_s = 0.891$, $p < .001$), validating the temporal and genre-stable nature of LVC frequency patterns. Hypothesis 2 confirms a significant deviation from Zipfian expectations, aligning more closely with Zipf–Mandelbrot law. Additional modeling supports the Menzerath–Altmann law (morpheme-based fit), while vocabulary dynamics are elaborated through Heaps' Law and Baayen's productivity metrics. Lexical drift is capture diachronically via Yule's K and KL Divergence, while entropy measures underscore shifting lexical concentration. A contrastive analysis with Altmann's (1967) seminal dictionary-based lexical model of Indonesian reveals substantial structural divergence in morpheme density, type distribution, and correlation behavior. Hypothesis 3 is confirmed through morpho-semantic-syntactic stratification across clusters, identifying a structurally asymmetric, gradiently layered LVC system. The secondary part classifies verb elements into True Light Verbs and Vague Action Verbs using aktionsart diagnostics. Machine learning (Naïve Bayes and Random Forest) highlights frequency and semantic parameters as strong predictors of verb productivity. The final part analyzes noun components, distinguishing stative and eventive types based on temporal features. Findings indicate that stative interpretations are largely noun-driven, while eventive reading emerge from verb-noun interaction and distributional patterns. This research offers a theoretically informed, data-driven typology of Indonesian LVCs, contributing to corpus linguistics and the broader modeling of LVC systems in underdescribed languages.

Absztrakt

A disszertáció az indonéz könnyűige-konstrukciókat (Light Verb Constructions, LVCs) elméleti nyelvészeti és korpuszalapú elemzés integrált megközelítésével vizsgálja. A kutatás elsődleges része empirikus alapot nyújt, négy különböző korpusz alapján értékeli az LVC-k gyakorisági és eloszlási mintázatait. A K-means klaszterezés három természetes csoportot azonosít (alacsony, közepes és magas gyakoriság), és az 1. hipotézist erős keresztkorpusz-rangsor-konzisztencia támasztja alá (Spearman $r_s = 0.891$, $p <.001$), ami megerősíti az LVC-k időbeli és műfaji stabilitását. A 2. hipotézis igazolja az eltérést a klasszikus Zipf-eloszlástól, és a Zipf–Mandelbrot-modellhez való közeledést. További modellezés alátámasztja a Menzerath–Altmann törvényt (morfémaalapú illeszkedés), míg a lexikai növekedést Heaps-törvény és Baayen-féle produktivitási mutatók részletezik. A lexikai eltérést időben Yule's K, KL Divergencia és entrópiaindexek vizsgálják. Altmann (1967) szótáralapú modelljével való összevetés jelentős eltéréseket tár fel a morfémahossz, a típuseloszlás és a korrelációs viszonyok terén. A 3. hipotézist a klaszterek közötti morfoszemantikai-szintaktikai rétegzettség támasztja alá, amely egy strukturálisan aszimmetrikus, fokozatosan rétegzett LVC-rendszert mutat. A másodlagos rész a verbális elemek osztályozására összpontosít, megkülönböztetve a Valódi Könnyű Igéket (True Light Verbs) és a Homályos Cselekvő Igéket (Vague Action Verbs) aktionsart diagnosztikák alapján. A gépi tanulási modellek (Naïve Bayes és Random Forest) kiemelik, hogy a gyakorisági és szemantikai paraméterek erős prédikátorai az igei produktivitásnak. A záró rész a névszói összetevők elemzését végzi el, statikus és eseményszerű típusok szétválasztásával időbeli jellemzők alapján. Az eredmények azt mutatják, hogy a statikus értelmezések főként névszóvezéreltek, míg az eseményszerű olvasatok az ige–névszó kölcsönhatásából és az eloszlási mintákból erednek. A kutatás elméletileg megalapozott, adatvezérelt tipológiát nyújt az indonéz LVC-k számára, hozzájárulva a korpusznyelvészethez és a kevéssé leírt nyelvek LVC-rendszereinek modellezéséhez.

Chapter 1

Introduction

1.0 Outline

The present study centers on examining the formation of light verbs in an agglutinative language, with a particular emphasis on constructions originating from Bahasa Indonesia (Indonesian).[1] Methodologically, the research critically analyzes these constructions through the application of theoretical linguistic approaches and corpus-based methods. This chapter provides a comprehensive introduction to the study, including the research's motivating factors (§1.1), the inquiries it addresses (§1.2), the theoretical framework it employs (§1.3), the previous research that guides it (§1.4), and an overview of the dissertation's structure (§1.5).

1.1 Motivation

An exploration to Light Verb Constructions (henceforth: LVCs) in Indonesian is somehow limited quantitatively and qualitatively.[2] While the phenomenon undoubtedly merits in-depth investigation, the available studies have thus far provided a limited scope of inquiry. This limitation is further compounded by the status of Indonesian as one of the world's many low-

[1] It is noteworthy that Indonesian boasts an impressive speaker count of almost 275 million individuals worldwide as of 2024. On November 16, 2023, this language earned the distinction of being the tenth official language at the UNESCO General Conferences (UNESCO. General Conference, 2023). This designation places Indonesian in esteemed language alongside nine other official UN languages: English, Arabic, Mandarin, French, Spanish, Russian, Hindi, Italian, and Portuguese. This contextual or extra lingual status of Indonesian, in a way, highlight the importance to choose the language as the object of the present study.

[2] A recent author's scientometric analysis (2024) indicates a near-complete absence of scholarly publications on LVCs in the Indonesian linguistic context within reputable databases such as Scopus and Web of Science. This dearth of research, to a certain degree, underscores the potential significance of the present study in contributing to the specific field of Indonesian linguistics and, more broadly, in enriching the existing corpus of LVC research.

resource languages (henceforth: LRLs). As noted by Singh (2008), Cieri *et al.* (2016), and Tszetkov (2017), LRLs are often characterized by being less studied, resource-scare, less computerized, less privileged, less commonly taught, or of low density. In the context of Natural Language Processing (hereafter: NLP), a field that experienced a major shift from rule-based to statistical-based techniques in the 1990s, LRLs like Indonesian face challenges as most of today's NLP research focuses on a small fraction (around 20) of the world's 7,000 languages, making it difficult to directly apply statistical methods due to data scarcity (Maxwelll-Smith *et al.*, 2022; Magueresse *et al.*, 2020; Sebastian *et al.*, 2022). This scarcity of resources, including annotated corpora and NLP tools, necessitates a research design that incorporates both corpus-based analysis and introspective linguistic insights from native speakers. Such a combined approach aims to maximize the information available for the study of LVCs in Indonesian, acknowledging the limitations inherent in LRL research. Consequently, a comprehensive understanding of complex linguistic phenomena such us LVCs within the Indonesian linguistic system remains elusive, underscoring the need for further exploration and systematic analysis to elucidate the full range of their morpho-syntactic and morpho-semantic properties.

Furthermore, to achieve the primary objective of the present study, the analysis will be grounded in the rigorous definition of LVCs as articulated by Mel'čuk (2022), as follows:

"Definition: support verbs also known as light verbs [$V_{(support)}$].

> Let there be a noun N whose meaning is a semantic predicate 'σ': N('σ'). Then:
> A verbal lexeme V is a support verb $V_{(support)}$ if and only if
>
> 1) V is used with an N('σ') such that this N('σ') is its DSynt-actant:
> V-I/II/… → N('σ');
> and
> 2) the meaning of the phrase V → N('σ') is the same as that of N('σ'), i.e., it is 'σ':
> 'V→ N('σ')' = 'N('σ')' = 'σ'."

(Mel'čuk, 2022: 4-5)

According to Mel'čuk's (2022) formal definition, the concept of light verbs centers on a specific semantic relationship between a verb and a noun within a verb-noun construction. This definition posits that a verb, designated as a support verb, functions in conjunction with a noun to essentially express the semantic predicate inherent in the noun itself. Formally, if we consider a noun N whose meaning constitutes a semantic predicate 'σ' (represented as N('σ')'), then a verbal lexeme V qualifies as a support verb under two crucial conditions. Firstly, V must be employed with N('σ')' as its direct syntactic actant (DSynt-actant), denoted as V-I/II/…→

N('σ')'. This signifies that the noun bearing the semantic predicate is a core argument of the verb. Secondly, and fundamentally, the meaning of the resulting verb-noun phrase N('σ')' must be equivalent to the meaning of the noun N('σ')' alone, effectively mirroring the semantic predicate 'σ'. This can be represented as 'VN('σ')' = 'N('σ')' = 'σ'. In essence, the light verb acts as a grammatical carrier, enabling the noun's semantic predicate to be expressed within a verbal structure, without contributing substantial independent semantic content of its own. This phenomenon is crucial in understanding the nuances of LVCs, where the verb's role is primarily grammatical rather than lexically substantive.

LVCs, by their nature, represent a linguistically intriguing phenomenon prevalent across diverse range of languages (*see* Example 1.1 – 1.11). Contrary to the conventional expectation that such constructions are confined to inflectional languages (Butt, 2010), LVCs have also been recognized in other type of languages, demonstrating their remarkable adaptability to varying morphological systems (e.g., Bygi *et al.*, 2018; Nagy T. *et al.*, 2020; Miyamoto & Kishimoto, 2016; Nugraha & Vincze, 2024; Xu *et al.*, 2022). This versatility equips LVCs with the capacity to conform to the grammatical structures of diverse linguistic systems while simultaneously providing speakers with a versatile tool for expressing a wide spectrum of meanings. In major cases, the nucleus of an LVC comprises a semantically bleached verb, often conveying a general meaning akin to '*do*' or '*make*' in English, in conjunction with another element, typically a noun, which conveys the specific meaning of the construction (cf., Giparaitė, 2015, 2024; Ronan, 2014, 2019; Ronan & Schneider, 2015).

(1.1) Indonesian (Nugraha, 2024: 104)
memberikan *kuliah*
give-TR.PRED lecture-OBJ.P
'give a lecture'

(1.2) Hungarian (Vincze, 2011: 283)
kutatást *folytat*
research-ACC do
'do research'

(1.3) Persian (Eshaghi & Karimi-Doostan, 2023: 79)
hormat *da:sht*
respect have-PST.3SG
'was respectable among people'

(1.4) Turkish (Özge *et al.*, 2022: 11)

yardim *etmek*
hand do
'to help'

Specifically, these examples (1.1 – 1.4) offer a glimpse into the cross linguistic variation of LVCs, showcasing how different languages utilize this structure to achieve similar semantic goals, albeit with unique grammatical nuances. According to Mel'čuk's (2022) formal definition, the LVCs is consisting of a support verb ($V_{(support)}$) and a noun with a semantic predicate 'σ', where the resulting VN phrase mirrors the meaning of N'σ' alone. In (1.1), Indonesian '*memberikan kuliah*' (to give a lecture), the verb '*memberikan*' (to give) functions as the $V_{(support)}$, providing the necessary grammatical scaffolding for the noun '*kuliah*' (lecture), which carries the core semantic predicate of 'lecturing'. The resulting phrase, '*memberikan kuliah*', effectively conveys the meaning 'to lecture', equivalent to the semantic content embedded within '*kuliah*' itself. This construction exemplifies how Indonesian employ light verb to create LVC, where the light verb contributes minimal semantic content beyond grammatical function, allowing the noun to take center stage. Thereafter, in Hungarian (1.2) '*kutatást folytat*' (do research), '*folytat*' (to do/carry out) serves as the $V_{(support)}$, while '*kutatást*' (research), in the accusative case, provides the semantic predicate. The phrase '*kutatást folytat*' translates to 'do research', effectively verbalizing the concept encapsulated within the noun '*kutatást*'. This construction demonstrates how Hungarian utilizes LVCs to express activities or processes associated with the noun, highlighting the flexibility of this structure in capturing a wide range of semantic nuances.

Thereafter, the Persian '*hormat da:sht*' as in (1.3) offers a slightly different perspective on LVCs. Here, '*da:sht*' (to have), in the past tense, functions as the $V_{(support)}$, while '*hormat*' (respect) carries the semantic weight. The resulting LVC, '*hormat da:st*', translates to 'was respectable', effectively verbalizing the state or quality denoted by '*hormat*'. This example showcases how LVCs can be used not only to express actions or processes but also to describe states or attributes, further expanding the functional range of this construction. Lastly, in Turkish (1.4) '*yardim etmek*' (do held/to help), the verb '*etmek*' (do) acts as the light verb, supporting the noun '*yardim*' (help). The construction '*yardim etmek*' expresses the act of helping, with the primary semantic contribution coming from '*yardim*'. The light verb '*etmek*' facilitates the integration of the noun into verbal phrase, enabling it to function as a predicate.

The semantic equivalence between the LVC and the noun's meaning aligns with formal definition.

Furthermore, the phenomenon of LVCs has been subject of extensive scholarly inquiry across diverse linguistic families (e.g., Avdeeva, 2017; Kettnerová, 2023; Vilas, 2017; Wang & Wu, 2020). Particularly, research on LVCs has conventionally been classified into three primary categories: investigations of isolating language, agglutinative languages, and fusional languages. Particularly in fusional (*see* Example 1.5–1.9) and isolating languages (*see* Example 1.10–1.11), LVCs have emerged as a focal point of linguistic analysis. The morphological traces left by LVCs have profound implications for the grammatical structure of a language (Buckingham, 2013; Dai, 2016; Vilas, 2015). These traces are not merely morphological relics but also convey essential semantic and syntactic information. In these languages with flexible word order, the construction of LVCs constitutes a distinctive linguistic manifestation that exhibits universal characteristics, such as transitivity and valency (Salido & Garcia, 2023; Wang *et al.*, 2023; Yim, 2020).

(1.5) English (Butt, 2010: 48)
'take a bath'

(1.6) Italian (Pompei & Piunno, 2023: 101)
prendere *una* *decisione*
take a decision
'take a decision'

(1.7) German (Fleischhauer, 2023: 373)
stehen *vor* *dem Bluten*
stand in_front_of the bleeding
'about to start bleeding'

(1.8) Spanish (Salido & Garcia, 2023: 248)
tomar *café*
take-INF coffee
'have a coffee'

(1.9) French (Mel’čuk, 2022: 4)

prendre *une* *decision*

take a decision

‘take a decision’

(1.10) Japanese (Miyamoto & Kishimoto, 2016: 425)

ryooko *o* *su-ru*

travel ACC do-PRS

‘takes a trip’

(1.11) Mandarin-Chinese (Xu *et al.*, 2022: 159)

Jin xing/zuo *le* *tiaozheng*

conduct/do COMPL adjustment

‘make an adjustment’

In detail, English, despite its predominantly analytic nature, also employs LVCs, as evidenced by the phrase ‘take a bath’ in (1.5). The verb ‘take’ acts as the $V_{(support)}$, while ‘bath’ provides the semantic predicate of ‘bathing’. The LVC ‘take a bath’ expresses the action of bathing, effectively mirroring the meaning of ‘bath’ in a dynamic, verbalized form. This example demonstrates how even languages with less complex morphology can utilize light verb to create LVC, highlighting the pervasiveness of this structure across different language types. Explicitly, in Italian example as in (1.6), the verb ‘*prendere*’ (take) functions as the light verb, supporting the noun ‘*decisione*’ (decision). The construction ‘*prendere una decisione*’ conveys the act of making a decision, with the core semantic content residing in the noun ‘*decisione*’. The light verb ‘*prendere*’ primarily contributes grammatical information, facilitating the integration of the noun into a verbal structure. The semantic equivalence between the LVC and the noun’s meaning aligns with the formal definition, as the phrase effectively verbalizes the concept inherent in the noun. Thereafter, the German examples, as in (1.7), presents an interesting case where the verb ‘*stehen*’ (stand) acts as light verb, combining with the prepositional phrase ‘*vor dem Bluten*’ (in front of the bleeding) to form an LVC. The construction expresses the state of being on the verge of bleeding, with the noun ‘*Bluten*’ (bleeding) providing the core semantic content. The light verb ‘*stehen*’, along with prepositional phrase, creates a metaphorical framing of the situation, depicting it as an imminent event.

In the Spanish example, as in (1.8), the verb '*tomar*' (take) serves as the light verb, supporting the noun '*cafë*' (coffee). The construction '*tomar cafë*' signifies the act of drinking coffee, with the semantic focus on the noun '*cafë*'. The light verb '*tomar*' primarily contributes grammatical information, while the core meaning resides in the noun. This example demonstrates how LVCs can express common activites using semantically light verb to verbalize a noun. Next, the French LVC as in (1.9) mirrors the Italian example, with '*prendre*' (take) acting as the light verb and '*decision*' (decision) providing the semantic core. The construction '*prendre une decision*' conveys the act of making-a-decision, with the light verb primarily contributing grammatical information. This cross-linguistic similarity highlights the prevalence of certain verbs, like 'take', in functioning as light verbs across different languages.

In some isolating languages, for instance Japanese as in (1.10) and Mandarin Chines as in (1.11), LVCs also manifest in certain way. The Japanese example showcases the light verb '*su*-ru' (do) combined with the noun '*ryooko*' (travel) and the accusative marker '*o*'. The construction '*ryookoo o su-ru*' expresses the act of traveling, with the semantic focus on the noun '*ryooko*'. This example demonstrates how LVCs in Japanese utilize particles to mark grammatical relations. Thereafter, in Mandarin Chinese LVC, the verbs '*jin xing*' (conduct) or '*zuo*' (do) function as light verbs, supporting the noun '*tiaozheng*' (adjustment). The construction '*jin xing/zuo le tiaozheng*' expresses the act of making an adjustment, with the core semantic content residing in the noun '*tiaozheng*'. The light verbs primarily contribute grammatical information and aspectual marking through the completive particle *le.* This example illustrates how LVCs in Chinese utilize word order and particles to convey specific meanings.

Moreover, agglutinative languages, characterized by their morphological complexity (cf. Haspelmath, 2009, 2012, 2020), present a particularly intriguing domain for the study of LVCs. When situated within a fixed-word order context such as Subject-Verb-Object (SVO), LVCs in agglutinative languages often exhibit distinctive morphological and semantic properties that are essential to an empirical understanding of these constructions. These constructions offer an exclusive way of combining a semantically light verb, which lacks inherent meaning, with noun element to convey specific meaning. As a matter of fact, LVCs are ubiquitous in many agglutinative languages, including Indonesian amongst others. For instance, in Indonesian, '*membawa*' (bring) is an exciting light verb due to its resourcefulness. It can act as a grammatical scaffold, seamlessly combining various complements to create semantic degrees. The verb's meaning depends on the complement it partners with, allowing it to take on a range of interpretations. For case, '*membawa pesan*' (bring a message) expresses a 'straightforward action of carrying something'. However, '*membawa kemenangan*' (bring a victory) takes on a more abstract meaning, implying the 'action to win'. '*Membawa*' can also be used with complements like '*perasaan*' as in '*membawa perasaan*' (lit. bring the feeling) to describe 'conveying an emotion' or with '*perubahan*' as in '*membawa perubahan*' (lit. bring the change) to 'signify introducing a new state'. This adaptability makes '*membawa*' an appreciated instrument in the Indonesian language, enabling the expression of a rich tapestry of ideas and actions, as well as the others form of LVCs.

Considering some of this background, especially the universal nature of LVCs in various languages, to support the aim of exploring LVCs in Indonesian, the corpus linguistics approach employed in the current study. It underscores the significance of empirical analysis in understanding those language phenomena in the language-specific patterns. By examining a large dataset of Indonesian LVCs, we were able to identify patterns and trends that would have been difficult to discern through introspection or intuition alone. This methodology provides a solid foundation for understanding the complex interplay between the various light verb and their accompanying nouns, revealing the multifaceted nature of these constructions and their role in conveying meaning within the Indonesian grammatical context.

1.2 Research questions

This study aims to investigate and elucidate the phenomena of LVCs in Indonesian using a theoretical linguistic approach that relies on the corpus technique. Admittedly, the following are some technical questions regarding LVCs in Indonesian that require in-depth analysis and explanation:

1) What are the distinct linguistic forms of LVCs in Indonesian, and how are they characterized in terms of frequency and distribution within a corpus-based analysis?
   a) How does the observed frequency of LVCs differ across hypothetical and genuine datasets within the four corpora under investigation?
   b) How is the distribution of Indonesian LVCs in morphological tests affected by the parameters of base morphological dimension and extended morphological area?

   c) How does the distribution of Indonesian LVCs pattern in semantic tests when considering the parameters of synonymous immensity and advanced semantic allotment?
   d) What is the distribution of Indonesian LVCs in syntactic tests when the parameters of basic transitivity range and valency scope are applied?
2) To what extent is the theoretical explanation for the results obtained from the corpus analysis, and how can they be applied to understand further the nature and characteristics of verb elements within Indonesian LVCs?
   a) How does the aktionsart of verbal elements in Indonesian LVCs serve as a foundation for verb classification?
   b) What is the frequency and distribution of verbal elements in Indonesian LVCs?
   c) How is the analysis of verbal element conducted based on a selection of machine learning algorithms?
3) What are the linguistic features and traits of LVCs in Indonesian make them distinct from other types of constructions by considering their noun elements?
   a) How does the aktionsart of nominal elements in Indonesian LVCs serve as the foundation for classifying nouns as the nucleus of such constructions?
   b) What are the types of nouns and to what extent do they manifest in Indonesian LVCs?
   c) How can the features of nouns as heads of Indonesian LVCs be identified and analyzed based on their interaction with verbal elements?

The research question articulated in this study are designed to provide a comprehensive understanding of LVCs within Indonesian. The rationale behind these inquiries stems from a desire to bridge theoretical linguistic frameworks with empirical corpus-based methodologies. The initial set of questions aims to establish a clear typology of Indonesian LVCs, examining their frequency and distribution across diverse corpora. This allows for the identification of patterns and variations in LVC usage, grounded in authentic language data. Subsequent questions delve into the morphological, semantic, and syntactic properties of LVCs, employing specific parameters to ensure rigorous and systematic analysis. Furthermore, the inquiry into the theoretical underpinnings of the findings seeks to integrate corpus-based insights with established linguistic principles, particularly concerning aktionsart and verb classification within LVCs. Finally, the focus on the nominal elements with LVCs addresses the crucial role of nouns in these constructions.

1.3 Theoretical frameworks

1.3.1 An overview of Indonesian word formation

The Indonesian language offers a compelling example of word formation through the agglutinative process. This concise overview delves into the fundamental mechanisms that shape Indonesian word formation, elucidating how morphemes (semantically meaningful units) are concatenated to produce novel words and convey various grammatical functions. The morphology of the Indonesian language has been a topic of significant scholarly examination for a long time. Early works by scholars such as Uhlenbeck (1952) classified Indonesian as a predominantly agglutinative language. Agglutinative languages are known for considerably using suffixes to modify root words and create new ones while maintaining a transparent meaning for each morpheme (Arad, 2005; Bauer, 2019; Bauer *et al.*, 2015). In Indonesian, for instance, the addition of the suffix {-kan} to a verb creates a causative form (e.g., '*baca*' (lit. read) to '*bacakan*' (lit. to cause someone to read). However, the addition of suffix {me(N)-} to verb creates an agentive form (e.g., '*membaca*').

Specifically, the Indonesian language has some fundamental processes for word formation, i.e., (a) affixation, (b) reduplication, and (c) compounding. Affixation is the primary process involving adding prefixes, suffixes, infixes, or circumfixes to the root words. This process helps modify the meaning of verbs and derive other categories, such as nouns, adjectives, and numerals, to form a verbal construction. The Indonesian language employs various affixes, such as {me(N)-} for active voice and {di-} for passive voice, to alter the verb's meaning and create new categories.[3] Additionally, infixes {-em-} and {-el-} are used for emphasis and intensification of the verb's meaning, whereas the causative suffix {kan-} generates new verbs from nouns and adjectives. The circumfix {ke-an} (active voice) is also used to modify the verb's meaning. In summary, the affixation process in Indonesian plays a vital function in word derivation.

Table 1.1 provides a list of Indonesian affixes and their example in the word formation. It provides a representative overview of the diverse array of affixes employed in Indonesian

[3] As outlined in the subsequent chapter on dataset development (*see* Chapter 2), the two datasets of LVCs analyzed in this study predominantly feature verbs marked with the {me(N)-} prefix. This prefix, a common morphological marker in Indonesian, serves to indicate the object of the verb, thereby highlighting the complement's role in the sentence structure. Sometimes, the complement's role is taken by the nucleus of LVCs, i.e. the noun element. The prevalence of {me(N-)} marked verbs in our datasets underscores the significance of its morphological feature in understanding the syntactic and semantic properties of LVCs in Indonesian.

word-formation, illustrating the language's rich morphology and the varied semantic and grammatical functions these affixes fulfill. The table is organized by affix type, namely prefixes, infixes, suffixes, and circumfixes, reflecting the position of the affix relative to the base word. For each affix, the table presents its form, the base word it attaches to, and an example of the derived word along with its English gloss. The prefix category showcases the variety of initial affixes that can be added to base words, often altering the word's meaning or grammatical category. For instance, the prefix *me(N)-* can attach to verbs like '*bawa*' (carry) to form '*membawa*' (to bring), indicating a transitive action.

Table 1.1: Sample of affixes and affixation in Indonesian word-formation.

No	Type of affix	Affix	Base	Examples
1.	Prefix	{me(N)-}	*bawa*	*membawa* 'to bring, to carry'
		{N-}	*kopi*	*ngopi* 'to drink a coffee' (colloquial)
		{di-}	*tulis*	*ditulis* 'written'
		{ber-}	*jalan*	*berjalan* 'to walk'
		{ter}	*jatuh*	*terjatuh* 'to fall'
		{pe(N)-}	*tinju*	*petinju* 'boxer'
2.	Infix	{-el-}	*gembung*	*gelembung* 'bubble'
		{-er-}	*gigi*	*gerigi* 'tooth' (as in a saw)
		{-em-}	*kuning*	*kemuning* 'a type of smell tree with fragrant flowers'
		{-in-}	*kasih*	*kinasih* 'beloved'
3.	Suffix	{-kan}	*buku*	*bukukan* 'to book (something)'
		{-i}	*bumbu*	*bumbui* 'to spice'
		{-in}	*bersih*	*bersihin* 'to clean'
		{-an}	*imbal*	*imbalan* 'reward'
		{-nya}	*diri*	*dirinya* 'himself/herself/itself'
		{-lah}	*kamu*	*kamulah* 'it is you'
		{-kah}	*bagaimana*	*bagaimanakah* 'how'
4.	Circumfix	{ber – an}	*dua*	*berduaan* 'as a pair'
		{ber – kan}	*dasar*	*berdasarkan* 'based on'
		{ke – an}	*besar*	*kebesaran* 'majesty'
		{pe(N) – an}	*tulis*	*penulisan* 'writing'
		{per – an}	*tanya*	*pertanyaan* 'question'
		{se – nya}	*lazim*	*selazimnya* 'as is customary'

Infixes, while less frequent than other affix types, are inserted within the base word, typically altering its meaning or adding a specific nuance. The table presents examples of infixes *-el-*, *-er-*, *-em-*, and *-in-*, each contributing a distinct semantic flavor to the base word. For example, the infix *-el-* in '*gelembung*' (buble) from '*gembung*' (inflated) suggests a rounded or bulging shape. Suffixes, attached at the end of the base word, often mark grammatical relations or add specific semantic nuances. The table includes examples of suffixes *-kan* (causative marker), *-i* (locative/benefactive marker), *-in* (repeated action marker), *-an* (nominalizing marker), *-nya* (possessive marker), *-lah* (particle emphasizing politeness), and -*kah* (interrogative particle). These suffixes demonstrate the diverse roles they play in Indonesian grammar, from deriving new verbs and nouns to marking possession and emphasis. Circumfixes, a combination of a prefix and a suffix, simultaneously attach to both ends of the base word, often creating words with complex semantic and grammatical functions. The table provides examples of circumfixes *ber-an* (plural/collective marker), *ber-kan* (causative/instrumental marker), *ke-an* (abstract noun marker), *pe(N)-an* (nominalizing marker), *per-an* (resultative/locative nominalizing marker), and *se-nya* (adverbial marker). The circumfixes illustrate the ability of Indonesian morphology to create complex word forms with nuanced meaning, often derived from simpler base words. This understanding of Indonesian affixation is crucial for comprehensive analysis of the language, particularly in the context of studying complex constructions such as LVCs, where the interplay between verbs and nouns is central to their interpretation. The information presented in this table serve as one of the foundational references for the subsequent investigation of LVCs within the broader framework of Indonesian morpho-syntax and morpho-semantics.

In addition to affixation, reduplication is a process in Indonesian where a root word is repeated partially or fully for various purposes. It can be used to emphasize, imitate sounds, or indicate plurality. For instance, '*jauh-jauh*' means 'too far'. Table 1.2 provides examples of reduplication types in Indonesian. Specifically, reduplication is crucial in adding expressiveness and nuance to word formation in the Indonesian language. According to Sneddon *et al.* (2010), Indonesian reduplications are full, partial, and imitative. Regarding the first type of reduplication, lexical items may undergo full reduplication, a morphological process involving the repetition of the entire word in its entirety. This phenomenon encompasses both simple words, or free bases, and complex words composes of one or more affixes attached to a base form. In addition, partial reduplication in Indonesian is exclusively applicable to consonant-initial bases. This morphological process involves the prefixation of a syllable composed of the base's initial consonant followed by the vowel [e]. Lastly, imitative reduplication involves the repetition of a word with slight variation between the two parts, often involving consonants or vowels.

Table 1.2 presents a comprehensive overview of reduplication processes in Indonesian word-formation, highlighting the diverse forms and semantic functions of this morphological phenomenon. Reduplication, a process where a word or part of word is repeated, is a productive strategy in Indonesian for deriving new words with a range of meanings. The table categorizes reduplication into three primary types: full reduplication, partial reduplication, and imitative reduplication, each exhibiting distinct characteristics and semantic contributions. Full reduplication, the most common type, involves the repetition of the entire base word, often resulting in plurals, intensifications, or distributed reference. The table illustrates this with examples such '*buku-buku*' (books) from '*buku*' (book), '*perubahan-perubahan*' (changes) from '*perubahan*' (change), and '*gula-gula*' (sweets) from '*gula*' (sugar). In these cases, the full reduplication can also denote intensification, as seen in '*kuda-kuda*' (easel) from '*kuda*' (horse), where the repetition evokes the image of horse-like structure. Additionally, it can express a variety of something, as in '*langit-langit*' (ceiling) from '*langit*' (sky), where the reduplication refers to the ceiling as a part of the sky. The example '*mata-mata*' (spy) from '*mata*' (eye) demonstrates how full reduplication can create a new word with a specialized meaning. Lastly, '*duduk-duduk*' (sit about) from '*duduk*' (sit) exemplifies the iterative or continuous nature of an action.

Table 1.2: Sample of reduplications in Indonesian word-formation.

No	Type of reduplication	Base	Examples
1.	Full reduplication	*buku* 'buku'	*buku-buku* 'books'
		perubahan 'change'	*perubahan-perubahan* 'changes'
		gula 'sugar'	*gula-gula* 'sweets'
		kuda 'horse'	*kuda-kuda* 'easel'
		langit 'sky'	*langit-langit* 'ceiling'
		mata 'eye'	*mata-mata* 'spy'
		duduk 'sit'	*duduk-duduk* 'sit about'
2.	Partial reduplication	*luhur* 'noble'	*leluhur* 'ancestor'
		tangga 'ladder'	*tetangga* 'neighbour'
		jaka 'bachelor'	*jejaka* 'bachelor'
		laki 'husband'	*lelaki* 'man'
		tamu 'guest'	*tetamu* 'guest'
		tapi 'but'	*tetapi* 'but'
3.	Imitative reduplication	*sayur* 'vegetable'	*sayur-mayur* 'vegetables'
		lauk 'side dish'	*lauk-pauk* 'side dishes'
		ramah 'friendly'	*ramah-tamah* 'hospitable and friendly activity'
		warna 'color'	*warna-warni* 'all kinds of colors'
		gerak 'movement'	*gerak-gerik* 'movements'

Partial reduplication, in contrast, involves the repetition of only a portion of the base word, typically the initial syllable or morpheme. This process often results in the derivation of nouns or adjectives with specific semantic nuances. The table showcases examples such as '*leluhur*' (ancestor) from '*luhur*' (noble), '*tetangga*' (neighbor) from '*tangga*' (ladder), '*jejaka*' (bachelor) from '*jaka*' (bachelor), '*lelaki*' (man) from '*laki*' (husband), and '*tetamu*' (guest) from '*tamu*' (guest). In these instances, the partial reduplication often conveys a sense of relatedness, similarity, or approximation to the base word's original meaning. Lastly, imitative reduplication, a less frequent but intriguing type, involves the repetition of a base word with a slight phonetic alteration, usually in the vowel sound. This process often creates words that are onomatopoeic or suggestive of sounds, manners, or a variety of items. The table includes examples such as '*sayur-mayur*' (vegetables) from '*sayur*' (vegetable), '*lauk-pauk*' (side dishes) from '*lauk*' (side dish), '*ramah-tamah*' (hospitable and friendly activity) from '*ramah*' (friendly), '*warna-warni*' (all kinds of colors) from '*warna*' (color), and '*gerak-gerik*' (movements) from '*gerak*' (movement). These examples illustrate how imitative reduplication can add a sense of variety, approximation, or emphasis to the base word's meaning, often creating a more expressive or evocative word. In conclusion, Table 1.2 provides an overview of the diverse forms and functions of reduplication in Indonesian word-formation. The examples presented demonstrate the productivity of this morphological process, highlighting its ability to create new words with a wide range of semantic nuances.

Table 1.3: Sample of compounding in Indonesian word-formation.

No	Type of compounding	Component 1	Component 2	Indonesian Compound
1.	Noun + Noun	*kereta* 'carriage'	*api* 'fire'	*kereta api* 'train'
		rumah 'house'	*sakit* 'sick'	*rumah sakit* 'hospital'
		kamar 'room'	*mandi* 'bath'	*kamar mandi* 'bath room'
2.	Verb + Noun	*makan* 'to eat'	*siang* 'noon'	*makan siang* 'to have lunch'
		tulis 'write'	*tangan* 'hand'	*tulis tangan* 'handwriting'
3.	Adjective + Noun	*putih* 'white'	*mutiara* 'pearl'	*putih mutiara* 'pearly white'
		merah 'red'	*delima* 'pomegranate'	*merah delima* 'ruby'
4.	Numeral + Classifier	*tiga* 'three'	*ekor* 'classifier for animals'	*tiga ekor* 'three animals'
		lima 'five'	*lembar* 'sheet'	*lima lembar* 'five sheets'
5.	Prepositional	*bawah* 'under'	*tanah* 'soil'	*bawah tanah* 'underground'
		bawah 'under'	*tangan* 'hand'	*bawah tangan* 'underhand'
6.	Verb + Verb	*makan* 'eat'	*makan* 'eat'	*makan-makan* 'to eat or dine together'
		masak 'cook'	*masak* 'cook'	*masak-masak* 'to cook or prepare food playfully'

Lastly, the Indonesian language is also known for its unique feature of compounding, where new words are created by combining existing words (Tadmor, 2018). There are various types of compounds found in the language, for instance noun-noun, verb-noun, adjective-noun, numeral-classifier, prepositional, and verb-verb compounds. Noun-noun compounds are the most common type, which combine two nouns to form a new concept. These are highly productive in Indonesian, often creating new nouns that refer to specific objects or entities. For instance, '*kereta api*' means 'train', where '*kereta*' originally signifies 'carriage' and '*api*' denotes 'fire', thus evoking the image of a fire-powered carriage. Similarly, '*kursi kayu*', meaning 'wooden chair,' combines '*kursi*' (chair) with '*kayu*' (wood), while '*meja besi*' (iron table) joins '*meja*' (table) and '*besi*' (iron), directly indicating the material composition of these objects. In contrast to these noun-noun compounds, verb-noun compounds combine a verb and a noun to create a new verb. These compounds typically create new verbs that express complex actions or activities. For example, '*makan siang*' means 'to have lunch', '*makan malam*' means 'to have dinner', and '*makan pagi*' means 'to have breakfast'. Adjective-noun compounds combine an adjective and a noun to create a new noun, such as '*luar biasa*' (extraordinary), '*besar hati*' (proud), and '*merah delima*' (ruby). These compounds can give rise to new nouns that describe qualities or attributes.

Additionally, numeral-classifier compounds combine a numeral and a classifier specific to the type of object being counted, such as '*tiga ekor*' meaning 'three animals', '*lima lembar*' meaning 'five sheets', and '*dua biji*' meaning 'two seeds'. These are essential for counting nouns in Indonesian. Prepositional compounds are formed by combining a preposition with a noun to form a new noun. Examples include '*bawah tanah*' meaning 'underground', '*bawah tangan*' meaning 'underhand', and '*atas angin*' meaning '*upwind*'. These compounds often function as adverbs or prepositions. Verb-verb compounds are relatively rare, where two verbs are combined to form a new verb, such as '*makan-makan*' meaning 'to eat or dine together', '*masak-masak*' meaning 'to cook or prepare food playfully', and '*lari-lari*' meaning 'to run around'. These compounds often express a sequence of actions or a combined action.

Above all, the compound of verb-noun is the closest classification into LVCs. It indicates a close relationship between verb-noun compounds and LVCs in Indonesian. While both involve a verb and a noun in close syntactic proximity, there are key distinctions and overlaps that warrant further exploration. Verb-noun compounds in Indonesian are lexical units formed by combining a verb and a noun, creating a single word with a distinct meaning. The meaning of the compound is often non-compositional, meaning it cannot be fully predicted from the meanings of its individual components. For example, '*makan siang*' (lit. 'eat noon') means 'to have lunch', which is not simply the sum of 'eat' and 'noon'. This non-compositionality and lexicalization distinguish verb-noun compounds from LVCs. LVCs, on the other hand, involve a verb, often semantically bleached, and a noun that carries the primary semantic content. The verb in an LVC functions more as a grammatical element, providing aspectual or Aktionsart information, while the noun contributes the core meaning. For example, '*memberikan kuliah*' means ‘to give a lecture’, where ‘*memberikan*’ (give) is the light verb and ‘*kuliah*’ (lecture) carries the semantic weight. There are also crucial differences. In LVCs, the noun typically retains its independent semantic status and can often stand alone as a noun phrase. In contrast, the noun in a verb-noun compound loses its independent status and becomes an integral part of the compound word.

1.3.2 Status and properties of LVCs

LVCs are a complex phenomenon found in numerous languages. They involve the combination of a lexical verb, often semantically weak, with another element, typically a noun or noun phrase that carries the core meaning of the construction. This section briefly overviews LVCs, including their grammatical function, semantic contribution, and theoretical implications. Table 1.4 presents a concise overview of alternative definitions for LVCs culled from various linguistic studies, categorized by their dominant perspective. The table highlights five primary status and properties: semantic, syntactic, functional, contrastive, and cross-linguistic. The semantic definition emphasized the semantic bleaching of the light verb, which acts as a support for the contentful noun, contributing minimally to the overall meaning. The syntactic definition focuses on the structural role of the light verb within the verb phrase, often likened to an auxiliary or a component of a complex predicate. Functionally, LVCs are viewed as multiword expressions serving diverse grammatical roles, such as marking aspect, causation, or modality. The contrastive definition distinguishes LVCs from full-verb constructions, highlighting the semantic lightness of the verb and its independence on a complement for a complete meaning. Lastly, the cross-linguistic definition acknowledges LVCs as a widespread phenomenon across languages, characterized by a light verb combined with a noun or adjective. Each explanation offers a unique lens through which to examine LVCs, reflecting varying theoretical frameworks and analytical foci. These diverse perspectives underscore the complexity of LVCs and the ongoing scholarly debate surrounding their precise characterization. This table provides a resource for navigating the varied literature on LVCs and understanding the nuances of their definition across different linguistic approaches.

Historically, the origins of LVC analysis can be traced back to the Renaissance period when notable scholars such as Elio Antonio de Nebrija (1444-1522) and Johann Christoph Adelung (1732-1806) made significant strides in classifying verbs based on their semantic significance. Interestingly, some verbs were identified as having weaker meanings and thus required further elements for complete predication (e.g., Coulmas, 2016; De Nebrija & Armillas-Tiseyra, 2016; Considine, 2014; Strohbach, 1984). These early observations served as a foundation for future discussions regarding LVCs. During the 18th century, there was a noticeable shift towards a more organized and systematic language analysis. Distinguished scholars such as Franz Bopp (1791-1867) and Hermann Paul (1846-1921) delved deeper into the concept of verb phrases, acknowledging the potential for auxiliary verbs to combine with main verbs and create complex predications (Chen, 2018; Frawley, 2003; Koerner, 2008; Reis, 1978). While these researchers recognized the existence of such complex linguistic constructions, explicit discussions of LVCs as a distinct phenomenon remained limited during that period.

The term "light verb" was initially coined by Otto Jespersen, a prominent Danish linguist, in his seminal work, *A Modern English Grammar on Historical Principles*, *Volume VI, Morpholgy*.[4] Jespersen identified certain verbs, such as "take," "do," and "get," as light verbs due to their semantically bleached nature and reliance on a subsequent noun or adjective for complete sense, as in "do the dishes" and "get a degree" (Jespersen, 1965; 2013). He posited that LVCs exemplified a language's ability to express intricate ideas through multiword units (Falk, 1992; Jespersen, 2003, 2015). Subsequently, linguists like Charles Fries (1952) and Zellig Harris (1957) adopted a structuralist approach, analyzing LVCs based on their distribution within a sentence and the syntactic patterns they formed (e.g., C. C. Fries, 1954, 1961; P. H. Fries, 2008; Harris, 1951, 1957, 1970). This approach laid the groundwork for further syntactic analyses of LVCs, advancing the field of linguistic theory and contributing to a deeper understanding of language structure and function. In the latter half of the 20th century, a surge of theoretical frameworks analyzed LVCs. Generative linguists, such as Emmon Bach in 1979, explored LVCs within verb phrase movement, proposing that light verbs move to higher syntactic positions within a sentence (e.g., Bach, 1965, 1988, 2005). Others, like Geoffrey Lakoff in 1977, focused on the semantic contribution of light verbs, suggesting that they contribute aspectual or causative meaning to a construction (Lakoff, 1993; Lakoff & Johnson, 1980). Recent decades studies, notably, delve deeper into the cross-linguistic variation of LVCs (e.g., Comrie, 1976, 2000, 2017).

Moreover, the grammatical role of LVCs is a subject of ongoing debate. Some contend that light verbs function as auxiliary verbs, supplying information about tense, aspect, or mood (for example, "She took a walk" versus "She is walking") (e.g., Caro & Arús-Hita, 2020; Yim,

[4] *See also* rerpinted version of other notable works by Jespersen (2010, 2013a, 2013b, 2013c, 2013d, 2015).

2020). Others consider them to be part of a single verb phrase, in which the light verb provides functional significance, such as causation ("He made a mistake") or completion ("We had a discussion") (Hellan, 2023; D. Wang *et al.*, 2023). In addition, LVCs can serve as phrasal verbs, with the light verb losing some of its original meaning and becoming idiomatic (for example, "He kicked the bucket") (e.g., Miyamoto & Kishimoto, 2016; A. Wang & Wu, 2020).

Table 1.4: Alternatives definition of LVCs.

No	Type of definition	Details	Further readings
1.	Semantic	LVCs are a type of multiword expression that consists of a semantically weak verb and a noun that carries the primary meaning of the phrase. The light verb in an LVC is a grammatical placeholder that does not contribute much to the expression's overall meaning but facilitates the expression of tense, aspect, or mood. This means that the light verb is supporting in the construction, while the noun is the main component that conveys the intended message. LVCs can be found in many languages and are often used to create more nuanced expressions that allow speakers to convey complex ideas or emotions with greater precision and subtlety.	(Brugman, 2001; Fleischhauer, 2021; Ježek, 2023; Mehl, 2019)
2.	Syntactic	LVCs are verb phrases that typically incorporate a light verb functioning as either an auxiliary verb or a component of a singular verb phrase, depending on the particular theoretical framework being employed. The light verb's functional significance is conveyed by contributing to the verb phrase. Some experts assert that light verbs operate in a manner similar to auxiliary verbs, while others maintain that they are fully integrated into the verb phrase itself.	(García-Pardo, 2021; Kettnerová, 2023; Nugraha, 2023c; Pompei, 2023a; Si, 2021)
3.	Functional	LVCs are multiword expressions consisting of a light verb and a complement, usually a noun. LVCs can serve various grammatical purposes, such as expressing aspects which refer to the speaker's perspective on the temporal structure of an event. They can also express causation, the relationship between an action and its effect. Furthermore, LVCs can be used to convey modality, which refers to the speaker's attitude towards the event, such as possibility or necessity. Lastly, LVCs are often used to form phrasal verbs, which have a different meaning than the sum of their parts.	(Barking *et al.*, 2022; Ong & Rahim, 2021; Piątkowski, 2019)
4.	Contrastive	LVCs are verb phrases that consist of a light verb and a noun or adjective complement. Unlike full-verb constructions, where the verb carries most of the semantic meaning, the light verb in LVCs carries only a minimal amount of meaning and requires additional elements to complete the sense of the phrase. In other words, the light verb alone does not convey a complete action or state and must be paired with a complement that provides more specific information about the nature of the action or state being described. LVCs are, therefore, distinct from single verbs, which can stand alone and express a complete meaning without requiring additional elements.	(Mel'čuk, 2022; Pompei, 2023b; Snider, 2021)
5.	Cross-linguistic	LVCs are a common linguistic phenomenon found in various languages, including English, Romance, and Germanic languages. LVCs are characterized by a light verb,	(Acedo-Matellán & Pineda, 2019; Mattissen, 2023;

No	Type of definition	Details	Further readings
		which carries little semantic content and is often combined with a noun or adjective to form a predicate. This construction expresses various meanings, such as causation, aspect, and result. For example, Romance languages, i.e., Spanish, use the verb '*hacer*' (to do, to make) for causative constructions, such as '*hacer una pregunta*' (to ask a question) or '*hacer reír*' (to make laugh). In contrast, Germanic languages. i.e., German, utilize the verb '*nehmen*' (to take) for inchoative constructions, as in '*Eine Dusche nehmen*' (to take a shower) or '*eine Entscheidung treffen*' (to make a decision). It is important to note that LVCs are not limited to these specific languages and can be found in many others.	Pompei & Piunno, 2023; Ricca, 2015)

Despite their relatively weak lexical content, light verbs play a critical role in shaping the overall semantics of a construction. They are capable of contributing nuanced aspectual information such as completion or ongoing action, as well as modality, including both obligation and possibility. The choice of light verbs can also affect the thematic roles assigned to the arguments within the construction, as seen in the difference between "take a nap" and "have a nap." Ultimately, it is clear that light verbs are not to be underestimated in their importance in constructing meaning in language (Butt, 2010; Fleischhauer, 2021; Fleischhauer & Gamerschlag, 2014; Ronan, 2019).

Furthermore, the presence of LVCs challenges conventional notions of verb phrase structure and the role of lexical semantics. Such constructions raise questions regarding the nature of lexical versus functional significance and the interplay between morphology and syntax. Some theoretical frameworks situate LVCs within the broader context of verb phrase movement or argument realization (e.g., Mastrofini, 2023; Mudraya *et al.*, 2008). In contrast, others approach them through lexical decomposition (e.g., Karimi-Doostan, 2005; Mel'čuk, 2022; Wittenberg & Piñango, 2011), in which the light verb imparts a specific functional meaning that combines with the noun to generate the overall interpretation.

1.3.3 Theoretical linguistics approaches to LVCs

1.3.3.1 Morphological foundations for LVCs analysis: Verb + Noun Compositionality

The morphological analysis of LVCs necessitates a foundational understanding of the linguistic unit. This construct, a fundamental concept in linguistic theory, serves as the building block for words and sentences. Within the context of LVCs, two primary principles govern the morphological considerations: the type of word element and its structural relationship within the LVC (c.f., Bauer *et al.*, 2013; Embick, 2013, 2015; Lieber, 2007). The first principle pertains to the nature of the word elements constituting LVCs. These elements can broadly be categorized into two types: base forms and affixed forms. Base forms represent the core lexical unit of a word, conveying the primary semantic content. In LVCs, base forms often correspond to verbs or nouns that serves as the nucleus or semantic head of the construction and the complementary element. For instance, in the English LVC 'take a walk', both 'take' and 'walk' are classified as the base forms. In addition, in the Indonesian LVC 'mengambil langkah' (take a step), element 'mengambil' is the affixed form and 'langkah' is the base form.

In addition, affixed forms are morphemes that are attached to base forms to modify their meaning or grammatical function. They can be prefixes, suffixes, or infixes. In LVCs, affixed forms can play a crucial role in indicating the complement's relationship to the verb, its case, or its thematic role. For instance, in the Indonesian '*memberi kesempatan*' (give an opportunity), element '*memberi*' is the affixed form. The affix 'me(N)-' marks the action of the verb in transitive manner. Furthermore, the second principle concerns the structural relationship between word elements within LVCs. Combining verbs and nouns is a lexical morphological process that creates a distinct meaning (cf., Andreou, 2017; Lieber, 2010, 2011; Lieber & Štekauer, 2011; Ricca, 2015). This linguistic phenomenon has been observed in various inflectional and agglutinative languages, leading to a comprehensive explanation of this construction's variations in form and meaning. The lexicalization process solidifies a syntactic phrase into a single word meaning, a crucial aspect of verb + noun composition. For instance, "give way" in English, originally a verb phrase, has become a multiword expression with a specific meaning. The resulting multiword expression often undergoes a semantic shift compared to the individual components. This can be metaphorical (Hüning & Schlücker, 2015), as in "kick the bucket" (die), or involve a narrowing of meaning, as in "download" (initially "to load down").

In some languages, verb + noun compositions involve morphological integration, such as the addition of affixes or changes to the verb stem. This further cement the compound's status as a single semantic unit (Ohnheiser, 2015). Verb + noun compositions can exhibit various head-marking patterns, where either the verb or the noun takes the primary role in determining the grammatical category of the compound; if the primary role is in noun, the construction entitled as the LVCs (Nagy T. *et al.*, 2020; Vincze *et al.*, 2013) In English, for example, “birdwatch” (noun) is formed from a verb, while “moonshot” (verb) is formed from a noun. While some verb + noun compositions are highly productive, allowing for the creation of new compounds based on existing patterns, others are more fixed and limited in their formation. Although Indonesian LVCs are not strictly lexicalized compounds, these constructions share some similarities. These compounds lose their literal meaning and take on new meanings depending on the accompanying noun. Some constructions involve the addition of particles or changes to the verb stem, suggesting a level of integration. The morphosemantic properties of LVCs are significantly determined by selecting their core noun element. This crucial component plays a vital role in determining the overall meaning and grammatical behavior of LVCs.

1.3.3.2 Semantics foundations for LVCs analysis

In linguistic terms, a *light verb* is a verb that has lost its original meaning to become more abstract (Fleischhauer *et al.*, 2019), serving mainly to fulfil grammatical functions (Fleischhauer, 2023). In order to embody specific semantic content, the light verb relies heavily on its complement, the core noun, in LVCs (Nugraha, 2023a, 2023b). This relationship between the light verb and the core noun has implications for the adaptation of the meaning of the core noun, resulting in new meanings that vary according to possible interpretations based on the grammatical environment (Fleischhauer & Neisani, 2020).

In general, the theoretical framework developed by Mel’čuk (2022) is crucial for the semantics of LVCs, as it emphasizes the existence of semantic relations between light verbs and their head elements (cf., Levin & Hovav, 2017; Hovav & Levin, 2015). It also highlights the role of conceptual metaphors in LVCs and the relationship between semantic verbs and event structures in expressing different types of events and participants. Levin and Hovav’s typology of LVCs, based on the semantic and syntactic functions of LVCs, provides a guide for the analysis of LVCs in the realm of semantic coverage and grammatical behavior. Meanwhile, Langacker’s notion of “constructional meaning” is the key to understanding how LVCs combine with core nouns to create a variety of possible LVC semantic meanings (Booij, 2016; Caballero & Inkelas, 2013; Langacker, 2014).

Moreover, it is important to mention *aktionsarten* ‘aktionsart’ as the inherent factor within LVCs. In the current author’s considerate, Aktionsart, broadly defined, refers to the inherent properties of a verb that convey information about the event or state it denotes (Comrie, 1993, 1997; Perdue, 2006). In Agee’s explanation (2018), Bache’s *Aktionsarten* are descriptors signifying to the practical appearances of a situation. They are both lexical and grammatical, frequently differentiated theoretically from context, but predominantly, mainly in English, presumed by the meanings associated with a word (cf., Agee, 2018; Bache, 1982; Hatav, 1989; Hinrichs, 1985). *Aktionsarten* designate a situation's process through time, but do not present a relative point in time, as tense does (Comrie, 1981; Waltke & O'Connor, 1990). Accordingly, it is important to consider definition as follows:

> "Aktionsart is the semantic distinction, whether lexically or formally, between durative and punctual situations; references to time, but not deictically – *aktionsarten* are references to the way a situation carries out in the "flow" of time."
> (Agee, 2018: 5)

Aktionsart, a crucial concept in aspectual semantics, denotes the inherent temporal properties of a situation, distinguishing between durative and punctual event. As Comrie (1976) highlighted, this distinction can be encoded lexically, through the choice of verbs or other temporal modifiers, or grammatically, via aspectual markers. Crucially, aktionsart focuses on the internal temporal structure of an event, independent of its relation to the moment of utterance. This differentiates it from tense, which locates an event in time relative to the speaker's perspective. Aktionsart instead addresses how a situation unfolds within its own temporal boundaries – its 'flow' of time. For instance, 'running' is durative, occupying a span of time, while 'jumping' can be punctual, occurring at a specific point in time. This inherent temporal profile is essential for understanding how situations are conceptualized and described linguistically.

Moreover, in his notable work Binnick (1991: 201) presents Noreen's taxonomy of aktionsarten as in Figure 1.1. The taxonomy bifurcates at the highest node, distinguishing between *uniform* and *intermittent* aktionsarten. Uniform aktionsarten are characterized by a singular, uninterrupted eventuality, further subdivided into *momentary* and *durative.* Momentary verbs denote punctual occurrences, often termed punctual, perfective, or aoristic, signifying events with negligible duration. Conversely, durative verbs, also known as cursive or imperfective, portray actions extending over temporal interval. This category is refined by the inclusion of *virtual* and *agential* subcategories. Virtual duratives, encompassing inchoative, decessive, and perdurative aspects, describe events approaching, ceasing, or persisting through a state, respectively. Agential duratives, while not explicitly elaborated in this figure, imply a volitional agent sustaining the action. In contrast, intermittent aktionsarten, branching into *frequentative, iterative,* and *intensive,* denote actions marked by repetition or heightened intensity. Frequentative verbs indicate repeated occurrences at regular intervals, while iterative verbs suggest repetition with potential interruptions. Intensive verbs, on the other hand, emphasize the vigor or magnitude of the action, irrespective of its uniformity or intermittence. This taxonomy provides a structured framework for analyzing the nuanced temporal semantics inherent in verbal element of LVCs.

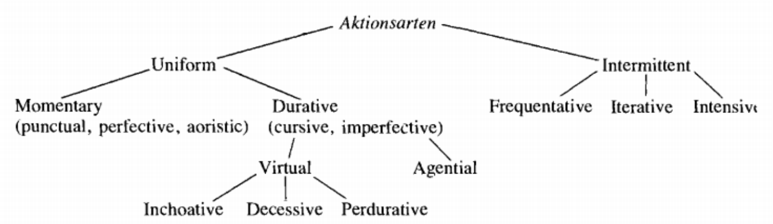


Figure 1.1: Noreen's taxonomy of aktionsarten (Binnick, 1991: 201).

In the context of LVCs, aktionsart plays a significant role in determining the compatibility of specific verbs with different event types, influencing the overall aspectual interpretation of the constructions. Analyzing aktionsart in Indonesian LVCs can reveal how these constructions encode and manipulate temporal information, contributing to a deeper understanding of their temporal function. One of the primary ways in which light verbs contribute to the aktionsart of LVCs is through their inherent semantic properties. For instance, the light verb '*membuat*' in Indonesian, which translates to 'make,' can be used to create LVCs with a variety of aktionsart. When combined with a noun like *pernyataan* 'statement' (e.g., make a statement), *membuat* can express a stative action. Meanwhile, when combined with *pertunjukan* 'performance' (e.g., make a performance), *membuat* can express an eventive action. In these cases, the light verb *membuat* provides the basic semantic framework for the LVC, while the specific aktionsart is determined by the lexical semantics of the noun and the contextual factors.

1.3.3.3 Syntactic foundations for LVCs analysis

Another important factor in determining the aktionsart of LVCs is the syntactic structure of the construction. The arrangement of elements within the LVC can influence the interpretation of the event or state being described. For instance, the presence of certain adverbial modifiers can change the aktionsart of an LVC. In English, the adverb *quickly* can modify the LVC *take a walk* to convey a durative state. Similarly, the use of a specific tense or aspect marker can also affect the aktionsart of an LVC. In Indonesian perfective aspect marker *sudah* 'already' can be used to indicate that an action has been completed, while the progressive aspect marker *sedang* 'currently' can be used to indicate that an action is ongoing.

Furthermore, the syntactic underpinnings of LVCs are also intertwined with notions of transitivity and valency (Arad, 2005; Baerman *et al.*, 2005; Blom, 2005; Embick & Noyer, 2007; Lieber, 2006; Roßdeutscher, 2014), both of which significantly influenced by the aktionsart of the construction. On the one hand, transitivity, a grammatical property that reflects the relationship between a verb and its arguments, is a pivotal factor in understanding LVCs. The aktionsart of an LVC can often predict whether it will be transitive or intransitive. Verbs that denote actions that involve a direct object tend to be transitive, while those that denote actions or states that do not require a direct object are typically intransitive. For instance, in Indonesian, the LVC '*memberi peluang'* (give an opportunity) is transitive, as the verb *memberi* requires a direct object (an opportunity). Conversely, the LVC '*jatuh hati*' (to fall in love) is intransitive, as the verb *jatuh* does not require a direct object within Indonesian grammatical context.

On the other hand, the aktionsart of an LVC can also influence its valency, which refers to the number and type of arguments a verb can take (Alexiadou *et al.*, 2015; Ausensi, 2021; Marantz, 2013). Verbs with a higher valency can take more arguments, while verbs with a lower valency can take fewer. In LVCs, the valency of the construction can be flexible or fixed. Flexible valency allows for the addition or omission of arguments without affecting the core meaning of LVC. For instance, the Indonesian LVC '*memberi komentar*' (give comments) is being used transitively or '*memberikan komentar*' (give comments to some beneficiary) is being used as complex transitive verb.

Lastly, the important syntactic framework is originated from Mel'čuk (2022) theoretical assumption as follows:

> "In other words, the meaning of the phrase $V_{(support)} \rightarrow N$ is identical to the meaning of full verb derivationally related to N and having the same meaning as N. Let us denote such a verb as V_0, N as S_0, and $V_{(support)}$ as $V_{(support)}(S_0(V_0))$. Then we can write:
>
> $$`V_0' = `S_0(V_0) \leftarrow V_{(support)}(S_0(V_0)'$$
>
> This semantic equality, expressed in term of lexical functions [LFs], underlies all paraphrastic manipulations with the $V_{(support)}$s."
>
> (Mel'čuk, 2022: 4)

Accordingly, there is a *nota bene* of the equation: the above equality holds only "ideally." In actual texts, it is often violated, just as beautiful physical laws are violated in ugly practical reality. This excerpt further clarifies the nature of support or light verbs by establishing a semantic equivalence between the verb-noun phrase (VN) and a corresponding full verb. It asserts that the meaning conveyed by the phrase $V_{(support)}$ + N is identical to that of a full verb V_0 derivationally related to the noun (N) and sharing its core meaning. This full verb, V_0, is conceived as the semantic equivalent of the noun, effectively encapsulating the predicate inherent in N. to formalize this relationship, the noun N is represented as S_0, and the support verb construction is denoted as $V_{(support)}(S_0(V_0)$. The semantic identity is then expressed as '$V_0 = S_0(V_0) \leftrightarrow V_{(support)}(S_0(V_0)$. This notation emphasizes that the support verb construction mirrors the semantic content of the full verb derived from the noun, highlighting the support verb's role as a vehicle for expressing the noun's inherent predicate rather than contributing independent semantic weight. Essentially, the interchangeability of the VN phrase with its corresponding full verb underscores the light verb's primary function as a grammatical support, devoid of substantial lexical autonomy.

1.3.4 Corpus-based method for LVCs analysis

Corpus-based methods offer a significant enhancement in analyzing LVCs. These methods involve leveraging extensive natural language text collections (corpora) (Biber, 1990; Egbert *et al.*, 2020), to obtain insights into the behavior and characteristics of LVCs used in real-world scenarios. Specifically, the corpus-based method or investigation is defined as follows:

> "The corpus-based investigation uses a corpus as a source of examples to check researcher intuition or to examine the frequency and/or plausibility of the language contained within a smaller data set."
>
> (Baker *et al.*, 2006: 49)

This excerpt elucidates the fundamental role of corpus-based investigation in linguistic research, particularly in validating theoretical assumptions and analyzing language patterns within a substantial body of authentic data. As Baker *et al.* (2006) suggest, a corpus serves as an invaluable resource for researchers, enabling them to empirically test their intuitions about language use. Rather than relying solely on subjective judgements, researchers can utilize a corpus to identify actual instances of the linguistic phenomena under scrutiny, thereby grounding their analysis in observable data. Furthermore, a corpus facilitates the examination of both the frequency and plausibility of specific linguistic structures or features. By quantifying the occurrence of particular patterns, researchers can gain insights into their prevalence and typical contexts of use. In context of studying Indonesian LVCs, a corpus-based approach is essential for identifying the range of verbs that function as light verbs, determining their associated semantic roles, and analyzing the syntactic environments in which they occur. This methodology ensures that the analysis is not solely based on theoretical constructs but is firmly rooted in the empirical reality of language use.

In like manner, McEnery & Hardie (2012) explains that:

> "Corpus-based studies typically use corpus data in order to explore a theory or hypothesis, typically one established in the current literature, in order to validate it, refute it or refine it. The definition of corpus linguistics as a *method* underpins this approach to the use of corpus data in linguistics."
>
> (McEnery & Hardie, 2012: 6)

McEnery & Hardie (2012) provide a succinct yet crucial characterization of corpus-based studies, articulating a perspective that is central to the methodological approach adopted in this research. They assert that corpus-based investigations typically leverage corpus data with the express purpose of exploring a pre-existing theory of hypothesis, often one that has been established within the current body of scholarly literature. This exploration serves a vital role in the research process: to validate, refute, or refine the theory or hypothesis under scrutiny. This methodological stance is of paramount importance in the context of analyzing LVCs. The use of corpus data allows for an empirical examination of theoretical claims about LVCs, moving beyond introspective judgements or limited sets of examples. By analyzing large amounts of naturally occurring language data, corpus-based methods provide a robust foundation for testing hypotheses about syntactic behavior, semantic properties, and distributional patterns of LVCs. The corpus serves as a testing ground where theoretical predictions can be confronted with real-world language use, enabling researchers to assess the validity and generalizability of existing theories. Furthermore, the corpus-based approach allows for the discovery of novel patterns or variations in LVC usage that may not have been anticipated by previous theoretical work, leading to a refinement or expansion of current understandings. In essence, McEnery and Hardie's description underscores the iterative relationship between theory and data in corpus linguistics, where empirical evidence from the corpus informs and shapes theoretical development.

Secondly, corpus data enables statistical analysis of LVC frequencies. This analysis can help identify the most common LVCs, their distribution across different grammatical conditions, and their potential changes in features. Lastly, corpus tools can be used to identify frequently occurring word combinations surrounding LVCs. Methodologically, the notion "frequency" within the present analysis is defined as follows:

> "The concept of frequency underpins much of the analytical work that is carried out within the remit of corpus linguistics. Frequencies can be given as raw data, e.g. there are 58,860 occurrences of the word 'man' in the British National Corpus (BNC); or (often more usefully) they can be given as percentages or proportions – 'man' occurs 602.91 times per million words in the BNC – allowing comparisons between corpora of different sizes to be made."
>
> (Baker *et al.*, 2006: 75)

This passage underscores the central role of frequency in corpus linguistics, highlighting its significance in quantitative analysis. As Baker *et al.* (2006) explain, frequency data can be expressed in two primary ways: as raw counts, representing the absolute number of occurrences of a given linguistic item, or as relative frequencies, such as percentages or proportions. While raw frequencies offer a basic measure of occurrence, relative frequencies are often more informative, particularly when comparing corpora of disparate size. Normalizing frequency data by calculating occurrences per a standard unit (e.g., per million words) allows for meaningful comparisons across different datasets, effectively mitigating the influence of corpus size. This standardization is crucial for revealing genuine differences in linguistic usage rather than merely reflecting variations in text length. In the context of Indonesian LVCs, frequency analysis can illuminate the prevalence of specific light verb + noun combinations, identify statistically significant patterns of usage, and track variations in light verb frequencies across different datasets. This quantitative approach, grounded in frequency data, is essential for uncovering statistically significant trends and patterns within the corpus. Overall, by systematically analyzing large corpora of language data, researcher can uncover significant trends, correlations, and statistical regularities that may not readily apparent through introspection or smaller-scale investigations. As suggested by McEnery and Hardie (2012: 51) that, "It is usually good practice to report *both* raw and normalized frequencies when writing up quantitative results from a corpus." Therefore, this corpus-based method can provide valuable insights into the typical roles associated with the LVCs within Indonesian grammatical context.

1.4 Previous notable studies

Previous notable research on LVCs has typically been cataloged into three primary categories: studies of agglutinative languages, fusional languages, and isolating language. The initial cluster of studies of LVCs in agglutinative languages encompasses Hungarian, Persian, and Turkish. Hungarian research exemplified by the works of Vincze (2011) has made significant contributions to this field. Similarly, Persian LVCs have been explored by Eshaghi and Karimi-Doostan (2023), while Turkish LVCs have been the subject of investigation by Özge *et al.* (2022), Rouhi and Heidari (2015), and Uçar (2010, 2012). These studies collectively provide a foundational understanding of LVCs in agglutinative languages, revealing commonalities such as the tendency for light verbs to grammaticalize from semantically fuller verbs and the significant role of case marking and agreement in LVC. Moreover, these investigations shed light on the diverse functions of LVCs, including their use in expressing aspectual distinctions, encoding causative relations, and altering argument structure. The insights gleaned from these studies on agglutinative languages provide a valuable insight for the current analysis of Indonesian LVCs.

Furthermore, the second group of studies focuses on fusional languages, including English (e.g., Kearns, 1998, 2002; Findlay, 2019; Giparaitė, 2016, 2023; Elenbaas, 2011; Fazly and Stevenson, 2007; Mehl, 2017; Bruening, 2015), German (e.g., Fleischhauer, 2021, 2023; Fleischhauer *et al.*, 2019; Fleischhauer and Gamerschlag, 2019; Hermann, 2019), Spanish (e.g., Salido and Garcia, 2023; Alonso Ramos, 2004, 2016; Bustos Plaza, 2005), French (e.g., (Mel'čuk, 2022; Abeillé & Godard, 2001) and Italian (Pompei *et al.*, 2023). These investigations delve into the intricacies of LVCs within languages characterized by their fusion of morphological elements, providing valuable insights into the cross-linguistic diversity of these constructions. Specifically, they illuminate the role of LVCs in encoding aspectual nuances, mediating argument structure alternations, and expressing complex predicates that involves abstract notions and events. These findings contribute significantly to the present study by offering a comparative perspective on functional and structural properties of LVCs in fusional languages.

Finally, the third group comprises languages known for their isolating morphology, including Chinese (e.g., Huang *et al.*, 2014; Huang & Lin, 2013; Xu *et al.*, 2022; Lin *et al.*, 2014; Cai *et al.*, 2019; Tsou & Yip, 2020; Dai, 2016), and Japanese (e.g., Miyamoto, 2000; Grimshaw & Mester, 1988). These languages, characterized by their relatively simple morphological structure, offer unique perspectives on the phenomenon of LVCs, particularly concerning the role of word order and syntactic context in their identification and interpretation. Research on Chinese and Japanese LVCs has revealed how these languages utilize particles, word order, and prosodic cues to distinguish LVCs from other verb-noun combinations. These findings provide valuable insights into the diverse strategies' languages employ to encode and signal grammatical relations in the absence of rich inflectional morphology. By considering these alternative strategies, the present study can gain a deeper understanding of the linguistic variation in LVCs and the interplay between morphology, syntax, and semantics in their formation.

1.5 Structure of the dissertation

This academic report is the culmination of the author's doctoral research and is divided into six chapters. Chapter 1 provides information on the motivation behind the research and problem formulation, as well as several theoretical frameworks and relevant previous research. Chapter 2 outlines the research methods, including information on materials or data, the corpus being analyzed, and procedures for collecting and analyzing data. Chapter 3 presents a comprehensive frequency and distribution analysis of LVCs in Indonesian. The frequency analysis involves the identification of LVC presence or absence across the two datasets within the four selected corpora. Additionally, this chapter offers a detailed examination of LVC distribution based on six empirical laws of language and six grammatical conditions embedded in six parameters. Chapter 4 delves into the structure of Indonesian LVCs, with a particular focus on the verb element. Chapter 5 extends this analysis by providing an in-depth examination of the noun element within Indonesian LVCs. Finally, Chapter 6 concludes the research, offering recommendations for future studies and acknowledging the limitations inherent in this investigation. Complementing all chapters, the appendix presents several important data and analysis results that are useful for understanding this research work comprehensively.

Chapter 2

Method of the study

2.0 Outline

This chapter presents a comprehensive exposition of the research methodology employed to investigate LVCs in Indonesian. The methodological approach is grounded in theoretical linguistics and strategically utilizes corpus linguistic techniques. Specifically, the investigation commences with a corpus-based analysis of LVCs, the empirical findings of which then inform subsequent theoretical inquiry. A detailed rationale for the methodological choices is articulated, encompassing the research design and approach (§2.1), materials utilized (§2.2), the procedure for data collection and analysis (§2.3), a review of pertinent prior author's work (§2.4), and ethical considerations (§2.5).

2.1 Research design and approach

The present study delved into the landscape of LVCs in Indonesian, a low-resource language, by employing a research design that integrated a corpus-based approach with principles from theoretical linguistics (*see* Figure 2.1). Recognizing the challenges posed by limited availability resources typical of LRLs, this integrated approach facilitated a comprehensive examination of LVCs, encompassing their frequency, distribution, underlying mechanisms, and linguistic contribution within the language (Biber, 2012; Egbert *et al.*, 2020). A large and representative corpus of Indonesian text constituted the core of the research design. This corpus provided the empirical foundation for the investigation, yielded a substantial amount of authentic language data in which LVCs could be observed in their natural context. Corpus analysis tools and techniques were employed to identify and categorize LVC types within the corpus. This process involved quantifying LVC occurrences and analyzing their distribution across different grammatical conditions. Additionally, the analysis also examined the inherent features of LVCs through their verb and noun elements. To some extent, this analysis yielded significant insights into the prevalence and preferred forms of LVCs in Indonesian.

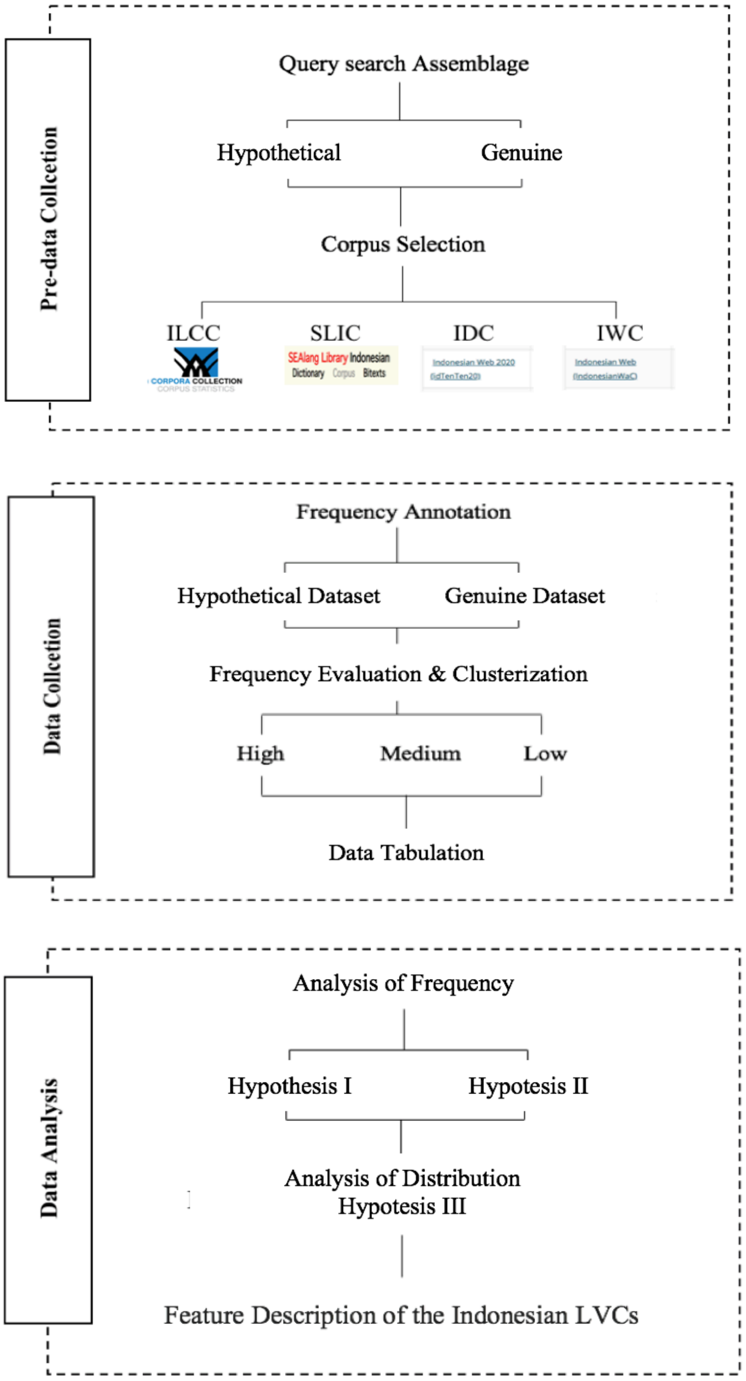


Figure 2.1: Schematic overview of the research process.

2.2 Materials

2.2.1 Data

The data for this study comprises Indonesian LVCs.[5] Concrete instantiations of these LVCs can be observed in examples (2.1) to (2.5). These instances, along with their frequency and underlying semantic and syntactic features, constitute the empirical data that form the research object of the current study. Accordingly, the Figure 2.2 presents an illustration on the total occurrence based on the raw frequency of Indonesian LVCs according to the two datasets within the four corpora in this study. In particular, the frequency data for each LVCs, within its initial categorization as *hypothetical* or *genuine*, obtained from four corpora in this study, serves as a crucial foundation for the quantitative analysis.

(2.1) *mengambil* *keputusan*
take-TR.PRED.RES decision-OBJ.P
'take a decision'

(2.2) *melakukan* *kajian*
do-TR.PRED.PROG study-OBJ.P
'do a study'

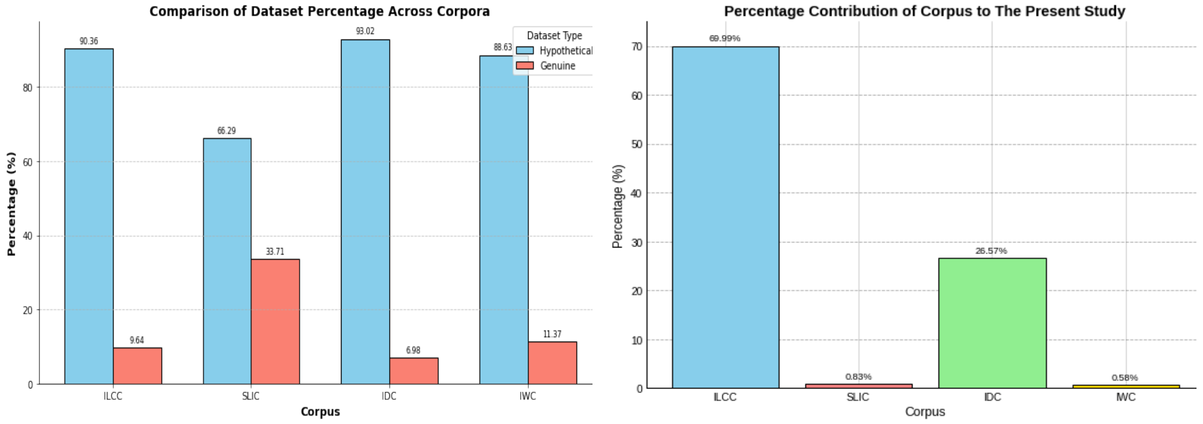
(2.3) *memberikan* *saran*
give-TR.PRED.PURP advice-OBJ.P
'give an advice'

(2.4) *membawa* *kebahagiaan*
bring-TR.PRED.REFL happiness-OBJ.P
'bring happiness'

(2.5) *mengajukan* *permintaan*
make-TR.PRED.PURP claim-OBJ.P
'make a claim'

[5] To overcome the lack of verifiability and reproducibility of the present study, the initial and processed frequency data are fully provided in the external repository, namely: https://github.com/dsnugrahaCL/LVCs_Indonesian.

In detail, Figure 2.2(a) provides a visual representation of the dataset percentage of LVCs across the four Indonesian corpora, i.e., the Indonesian-Leipzig Corpora Collection (ILCC), the SEAlang Library Indonesian Text Corpus (SLIC), the IdTenTen Indonesian Corpus (IDC), and the IndonesianWaC (IWC). The provided summary also details the internal proportional allocation between two distinct datasets, i.e., hypothetical and genuine. For ILCC, 90.36% originates from hypothetical dataset, with only a small fraction, 9.64%, stemming from genuine dataset. A similar trend is observed for IDC and IWC. IDC exhibits the highest dominance of hypothetical at 93.02%, leaving only about 6.98% for genuine dataset. IWC also shows a string leaning towards hypothetical dataset at 88.63%, with genuine dataset contributing around 11.37%. In contrast, SLIC presents a more balanced allocation between the two datasets. While hypothetical dataset still holds the larger share at 66.29%, genuine dataset constitutes a higher proportion compared to the other corpus items, making up 33.71% of its total data. This variation in proportional representation across the different corpus items highlights potential underlying differences in the nature and characteristics of these datasets within each specific context.



(a) (b)

Figure 2.2: Data proportion in four selected corpora. Part (a) displays proportion of the total LVC occurrences. Part (b) presents contribution comparison of corpora used for data retrieval.

In addition, Figure 2.2(b) presents the proportional representation of four distinct corpora – ILCC, SLIC, IDC, and IWC – within the overall dataset. The y-axis quantifies the percentage contribution, ranging from 0% to approximately 70%, while the x-axis delineates the individual corpora. The most significant contributor to the present study is the ILCC corpus, accounting for a substantial 68.99% of the total frequency. This dominance indicates that the majority of the frequency-data analyzed originates from this source, suggesting it holds the most weight in any subsequent analysis or modeling. In stark contrast, the SLIC and IWC corpora exhibit minimal contributions, at 0.83% and 0.58% respectively. Their small proportions imply that these datasets represent a relatively minor fraction of the overall study. Lastly, the IDC corpus occupies an intermediate position, contributing 26.57% to the total frequency. While significantly smaller than ILCC, its contribution is substantially larger than both SLIC and IWC, suggesting a notable presence within the analyzed dataset. The varying heights of the bars clearly illustrate the imbalanced distribution of frequency-data across the four corpora, highlighting the need to consider these differences in scale when interpreting any findings derived from this study. The visual disparity underscores the potential influence of the ILCC and, to a lesser extent, the IDC corpora on the overall results.

In this case, the analysis exemplifies the fundamental role of statistical analysis in corpus linguistics. As McEnery and Hardie (2012: 49) emphasizes, “Frequency data is so regularly produced in corpus analysis that it is rare indeed to see a study in corpus linguistics which does not undertake some form of statistical analysis.” The very act of normalizing our raw frequency data to pmw reflects this principle. By converting the raw counts into a standardized rate, we enable meaningful comparisons across corpora of varying sizes. This descriptive statistical approach, although basic, is indispensable for interpreting and drawing conclusions from large datasets. “To put it another way, any empirically based approach to linguistics which deals with large collections of data points may have cause to employ statistical analysis” (McEnery & Hardie; 2012: 49). Corpus linguistics, being inherently empirical, relies heavily on such analyses to reveal patterns and trends in language use. The observed variations in LVC frequencies across the four corpora, as depicted in the Figure 2.2(a), would be considerably less insightful without this essential statistical treatment. Therefore, the presented data, normalized and visualized, served as a demonstration to the integral role of statistical analysis in the present study.

2.2.2 Selected corpus as the data sources

As has been discussed in limited detail in the previous section, the present study utilizes data from four primary corpora to gather comprehensive information on LVCs. As a further explanation, statistical detail on these four corpora can be found in Table 2.1.

Table 2.1: Details identification of the selected corpus.

No	Aspect	ILCC	SLIC	IDC	IWC
1.	Descrip-tion	The Indonesian – Leipzig Corpora Collection is a mixed corpus based on material 2013. The majority of text comes from internet websites.	This monolingual corpus consists of Indonesian texts retrieved from a variety of internet sources.	The Indonesian Web Corpus (idTenTen) is an Indonesian corpus made up of texts collected from the internet.	The Indonesian web corpus (idWaC) is an Indonesian corpus made up of texts collected from the Internet.
2.	Name	Ind_mixed_2013	SEAlang Library Indonesian Text Corpus	idTenTen	idWaC
3.	Language	Indonesian	Indonesian	Indonesian	Indonesian
4.	Genre	Text-based (Online)	Text-based (Online)	Text-based (Online)	Text-based (Online)
5.	Year	2013	2010	2020	2012
6.	Sentences	74,329,815	2,242,565	258,435,545	6,003,769
7.	Types	7,964,109	5,000,000	3,678,192,045	90,120,046
8.	Tokens	1,206,281,985	15,763,657	4,432,864,160	109,236,814
9.	Token/sent.	16.23	7.03	17.15	18.19
10.	Link to corpus	https://corpora.uni-leipzig.de/?corpusId=ind_mixed_2013	http://sealang.net/indonesia/corpus.htm	https://www.sketchengine.eu/idtenten-indonesian-corpus/	https://www.sketchengine.eu/indonesianwac-corpus/

The selection of data sources for this research was predicated upon several critical criteria. These criteria encompassed: (i) the sources' established reputation for credibility and rigorous data collection methodologies, (ii) the direct relevance of their data to the research topic under investigation, and (iii) the comprehensive coverage they provided of the subject area. The selected sources underwent a rigorous evaluation process, focusing on data quality, accessibility, and alignment with the project timeline. Following careful deliberation, ILCC, SLIC, IDC, and IWC were chosen as the data sources, as they demonstrated optimal performance across these evaluative parameters.

Given the centrality of the corpus data to this investigation, the following part will provide a detailed description of the four corpora employed in this study. These corpora, carefully selected for their relevance and comprehensiveness, provide a rich resource of linguistic data for the analysis of Indonesian LVCs. ILCC, SLIC, IDC, and IWC each offer empirical perspectives on Indonesian language use. By drawing upon these varied resources, this study targets to capture a wide range of LVCs and gain a nuanced distribution within the Indonesian language. The subsequent description will delve into the specific characteristics of each corpus, highlighting their strength and limitations in relation to the research objectives of this dissertation.

2.2.2.1 Indonesian-Leipzig Corpora Collection (ILCC)

The ILCC is a part of the larger corpus, namely the Leipzig Corpora Collection which was created by Goldhahn *et al.* (2012) under the *Deutscher Wortschatz* project.[6] It is a monolingual corpus and a valuable resource for linguistic research on Indonesian, offering a substantial collection of texts compiled from diverse online sources. The ILCC comprises texts sourced from online newspaper, generic web pages, and Wikipedia, ensuring a wide range of topics and genres. This diversity is crucial for investigating the distribution and variation of linguistic phenomena, such as LVCs, across different contexts. The inclusion of newspaper texts provides access to formal written language, while web pages and Wikipedia offer examples of more informal and diverse writing styles. The ILCC has undergone several processing steps, including HTML stripping, language identification, sentence segmentation, and cleaning, to ensure data quality and consistency. While the developer does not explicitly mention annotation, the corpora have undergone statistical analysis to generate frequency information, co-occurrence data, and semantic maps for each word. This statistical information can be considered a form of annotation, as suggested by McEnery and Hardie (2012: 14) that "corpora which have already been analyzed in some way are annotated," providing valuable comprehensions into word behavior and relationships with the corpus. The ILCC's diverse text types, coupled with its statistical processing, make it a suitable resource for investigating the frequency, distribution, and semantic properties of LVCs in Indonesian.

2.2.2.2 SEAlang Library Indonesian Text Corpus (SLIC)

The SLIC stands as a monolingual corpus of Indonesian textual data, meticulously curated from a diverse array of internet sources. It is a product of a collaborative effort between CRCL and the University of Wisconsin-Madison Center for Southeast Asian Studies (CSEAS) (2010).[7]

[6] The *Deutscher Wortschatz* project, according to its official website of the corpora collection, has been delivering information about the German language since the mid-1990s by systematically collecting and processing available documents from the internet, typically annually. This effort has resulted in a corpus-based dictionary with a dedicated page for each word, including statistical information, example sentences, and links to related terms. With a wealth of text material encompassing several million sentences, the project can provide comprehensive insights into nearly every word, establishing it as one of the most extensive resources for the German language. Over the years, the service has expanded to include a broader range of languages under the Leipzig Corpora Collection, now offering for more than 250 languages that can be queried online, many of which provide some of the most freely available text resources, including Indonesian.

[7] The SEAlang Library was founded in 2005 to offer language reference resources for Southeast Asia, initially targeting the non-roman script languages of the mainland and currently focusing on the languages of insular

Unlike corpora built primarily from specific sources like newspapers or novels, the SLIC embraces a broader approach, encompassing a wide spectrum of text types, including web pages, blog posts, and online articles. This heterogeneity in source material ensures a representation of language use, capturing both formal and informal registers, diverse genres, and a multitude of topics. Furthermore, while the exact details of the annotation are not explicitly mentioned, the corpus has undergone meticulous processing. This preparation enhances the analytical potential of the corpus, enabling researchers to delve into grammatical patterns of a certain constructions. According to McEnery and Hardie's (2012: 14) characterization of corpora, SLIC can be classified as annotated. Specifically, the SLIC facilitates the investigation of collocational trends, providing insights into how words co-occur and form meaningful relationships within the text. This feature is particularly relevant to the study of LVCs, as it allows for the examination of the co-occurrence patterns between light verbs and their nominal complements. The SLIC's diverse composition makes it a well-suited resource for investigating the nature of LVCs in Indonesian.

2.2.2.3 IdTenTen Indonesian Corpus (IDC)

The IdTenTen Indonesian Corpus (IDC), provided under the Sketch Engine (Kilgarriff *et al.*, 2004; 2014)[8], distinguishes itself from the previously discussed corpora through its size and recency. Comprising a massive 3.6 billion words collected in 2020, the IDC offers a vast and up-to-date snapshot of Indonesian language use on the web. This extensive dataset provides a rich resource for investigating a wide range of linguistic phenomena, including the less frequent or nuanced aspects of language, which might be under-represented in smaller corpora. The

Southeast Asia. Its offerings encompass bilingual and monolingual dictionaries, monolingual text collections, aligned bitext corpora, and an array of tools designed for manipulating, searching, and displaying complex scripts. This initiative is a collaborative project between CRCL and the University of Wisconsin-Madison's Center for Southeast Asian Studies (CSEAS), which works closely with the Southeast Asian Studies Summer Institute (SEASSI) program, represented by fourteen-member institutions, including all US National Resource Centers for Southeast Asian Studies. Major funding is provided by the U.S. Department of Education's Technical Innovation and Cooperation for Foreign Information Access (TICFIA) program, complemented by matching funds from CRCL.

[8] Developed by Lexical Computing in 2003, Sketch Engine is a versatile corpus management and text analysis software designed for linguistic research. Its primary function is to facilitate the exploration of large text collections through complex, linguistically-motivated queries, enabling researchers to uncover patterns and insights in language data. Sketch Engine has been widely adopted in the field of corpus linguistics and is available in 11 languages, including Indonesian. The software is written in Python, Go, JavaScript, and C++, and was initially conceived by Adam Kilgarriff and Pavel Rychly. Sketch Engine's capabilities extend beyond basic corpus management, offering a range of features for analyzing word frequencies, collocations, concordances, and other linguistic phenomena.

IDC's size also enhances its statistical dominance, enabling more robust analyses and the identification of subtle patterns in language use. Furthermore, the IDC's focus on web data ensures that it captures the dynamic and evolving nature of online Indonesian, including emerging trends in vocabulary, grammar, and style. Beyond its size and recency, the IDC boasts a sophisticated annotation schema. According to McEnery and Hardie's (2012) characterization of corpora, IDC can also be classified as annotated corpora. In addition to part-of-speech tagging, which identifies the grammatical function of each word, the IDC also incorporates lemmatization. Lemmatization links each word form to its base form or lemma, enabling researchers to analyze words based on their core meaning rather than their surface form. This is particularly useful for studying LVCs, as it allows for the identification and analysis of different light verb.

2.2.2.4 IndonesianWaC (IWC)

The IndonesianWaC (IWC) is a massive web corpus of Indonesian, also offered in the Sketch Engine, comprising 100 million words harvested from the internet. Unlike the previously described corpora, which include texts from various sources like newspapers and Wikipedia, the IWC focuses exclusively on web-based content. This focus provides a snapshot of contemporary Indonesian language use as reflected in online communication, capturing a wide range of genres, registers, and styles prevalent on the internet. The IWC's size and web-centric nature make it a valuable resource for investigating linguistic phenomena that are particularly salient in online discourse, such as colloquialisms, internet slang, and emerging trends in language use. In term of annotation, the IWC has undergone part-of-speech tagging, a process that assigns grammatical labels (e.g., noun, verb, adjective) to teach word in the corpus. This annotation facilitates syntactic analysis and allows for the exploration of grammatical patterns within the data. Moreover, the IWC is equipped with a suite tools designed for corpus exploration and analysis. These tools enable researchers to generate word sketches, identify collocations, explore synonyms and related terms, extract keywords, and analyze n-grams, among other functions. For the study of LVCs, the IWC's focus on web-based Indonesian and its part-of-speech annotation, as well as the n-grams tool, offer valuable resources for examining the frequency and distribution of these construction within a contemporary Indonesian.

2.2.3 The unit of analysis

To address the central research questions regarding LVCs in Indonesian, this study adopts a corpus-based technique, necessitating a clearly defined unit of analysis. Given that these questions pertain to typology, theoretical underpinnings, and distinguishing features, the LVC itself constitutes the primary unit of analysis (*see* Example 2.6 – 2.8). Specifically, for the first research question, which explores LVC frequency and distribution, the unit of analysis are individual LVC occurrences within the all corpora. A sample of the unit of analysis for the first research questions is presented in Table 2.2.

Table 2.2: Sample of data-unit for frequency analysis.

No.	LVC	Total (Σ-Frequency)	Total (Z-Score)	PMW	Rank	Cluster
1.	*memberikan kuliah* 'to give a lecture'	12,532	-0.195	628.989	308	1
2.	*mengambil kesempatan* 'take a chance'	24,931	0.086	1252.203	206	2
3.	*membuat keputusan* 'make a decision'	114,153	2.106	5733.532	40	3

The second and third research questions are concerned with an investigation of the linguistic nature of LVCs. Here, the unit of analysis is broadened to consider not just the LVC as a composite structure, but also its inherent component parts, namely verb and noun elements. This refined approach enables a detailed exploration of how aktionsart affects the characteristics *inherent in* Indonesian LVCs. By adopting this expanded unit of analysis, which foregrounds the pivotal role of the verb and noun, the study presents an extended examination of their internal structure.

(2.6) *Terutama dipakai oleh kita dalam* ***mengambil keputusan*** *praktis.*
especially-ADV use-PASS by-ADP us-3PL in take-TR. decision-OBJ.P practical
'We primarily use this in making practical decision.'

(2.7) *Sebab pihaknya masih harus* ***melakukan kajian*** *dan analisa kasus.*
because their party still have do-TR. study-OBJ.P and analysis case
'That's because they still need to conduct a study and analysis of the case.'

(2.8) *Umar* ***memberikan saran*** *kepada Hz.*
Umar-PROPN.A.SBJ.TOP give-TR advice-OBJ.P to-ADP Hz-PROPN.BEN.SG
'Umar gave advice to Hz.'

Furthermore, examples (2.6) through (2.8) illustrate the LVCs that serve as the units of analysis in this study. Each example showcases an Indonesian LVC extracted from one of the corpora, highlighting the specific verb-noun combination that constitutes the construction. In example (2.6), the LVC *'mengambil keputusan'* (take decision) is presented, with '*mengambil*' (take) acting as the light verb and *'keputusan'* (decision) as the noun contributing the core semantic content. The construction can be translated as 'to take a decision' in relation to a sense of 'making a decision' in English version, demonstrating how the light verb grammaticalizes the noun, enabling it to function as predicate. Similarly, example (2.7) presents the LVC *'melakukan kajian'* (do study), where '*melakukan*' (do) serves as the light verb and '*kajian*' (study) provides the primary semantic meaning. This construction conveys the act of conducting a study, with the light verb facilitating the integration of the noun into a verbal phrase. Lastly, example (2.8) showcases the LVC *'memberikan saran'* (give advice), where *'memberikan'* (give) functions as the light verb and *'saran'* (advice) carries the core semantic weight. These examples collectively demonstrate the diversity of LVCs in Indonesian, highlighting the various verb-noun combinations that can form these constructions.

2.3 Procedures

This section delineates the procedures undertaken to conduct the current research, encompassing data collection and analysis. It outlines the steps involved in extracting relevant data from the four selected corpora, ensuring systematic approach. The data analysis procedures are subsequently detailed, specifying the methods employed to examine the properties and features of Indonesian LVCs.

2.3.1 Data collecting procedures

2.3.1.1 Query search assemblage

The data collection process comprises multiple stages, and the first stage involves preparing the data collection instruments.[9] In this case, the preparation involved using two attested forms of

[9] The complete instrument form of hypothetical and genuine matrices can be seen in the following repository: https://github.com/dsnugrahaCL/LVCs_Indonesian.

matrices: hypothetical[10] and genuine[11] matrices. The hypothetical matrix is a compilation of light verbs from literature studies of agglutination, fusion, and isolation languages used in prior research beyond the Indonesian language. This list can be considered adopted, meaning it is assumed to have not originated in Indonesian and is, to some extent, anticipated also to be found in Indonesian. The total initial-form of LVCs within hypothetical dataset is 900 items. On the other hand, the genuine matrix is a compilation of LVCs obtained through an introspective process with native Indonesian speakers. Although the genuine matrix has yet to be tested, it was compiled based on intuitive search results from Indonesian speakers with advanced knowledge and skills. The initial-form of LVCs within genuine dataset is 102 items. This initial assembly will enter the inclusion and exclusion stages. These stages are carried out based on specific parameters as described in the section 2.3.1.2.

Furthermore, it is important to explain the considerations for using the hypothetical and genuine datasets. The utilization of both hypothetical and genuine datasets in this research is underpinned by several key theoretical considerations. Drawing on Jespersen's (1924) assertion that any linguistic phenomenon can be regarded either "from without or from within," the present study strategically employs these datasets to capture a holistic view of Indonesian LVCs. The hypothetical dataset compiled from light verbs identified in prior research across various language morphological typologies (agglutinative, fusion, and isolating), represent the "from without" perspective. It provides an outward, externally validated starting point, allowing for the examination of cross-linguistic patterns of LVCs and investigating whether patterns observed in other languages are also prevalent in Indonesian. This approach acknowledges the potential for universal tendencies while recognizing language-specific variations. Conversely,

[10] The hypothetical matrix is a comprehensive collection of adopted light verb constructions derived from three distinct morphological typologies of language – agglutinating, fusional, and isolating. The first group comprises Hungarian (Vincze, 2011; Racz *et al.*, 2014), Persian (Family, 2014; Eshaghi *et al.*, 2023; Karimi-Doostan, 2023) and Turkish (Özge *et al.*, 2022; Uçar, 2010; 2012), while the second encompasses English (Kearns, 1998, 2002; Findlay, 2019; Giparaitė, 2016, 2023; Elenbaas, 2011; Fazly and Stevenson; 2007; Mehl, 2017; Bruening, 2015), German (Fleischhauer, 2021, 2023; Fleischhauer *et al.*, 2019; Fleischhauer and Gamerschlag, 2019; Hermann, 2019), Spanish (Salido and Garcia, 2023; Alonso Ramos, 2004, 2016; Bustos Plaza, 2005), and French (Mel'čuk, 2022; Abeillé *et al.*, 1998), and Italian (Pompei *et al.*, 2023). The third group includes Mandarin Chinese (Huang *et al.*, 2014; Huang & Lin, 2013; Xu *et al.*, 2022; Lin *et al.,* 2014; Cai *et al.*, 2019; Tsou & Yip, 2020; Dai, 2016), and Japanese (Miyamoto, 2000; Miyamoto *et al.*, 2016; Grimshaw & Mester, 1988). Each of these groups underwent a meticulous human-supervised translation process from the source language into Indonesian by following Mastrofini's (2023) technique and Nagy T. *et al.*'s (2020) general characteristics of LVCs across languages, culminating in the alignment of construction forms to ensure semantic coherence.

[11] The genuine matrix is a compilation of light verb constructions originating in Indonesian. The gathering of this formation was meticulous and involved introspection by native Indonesian speakers. This introspection stage allowed for a reflection on the cognitive-comprehensive framework of native speakers, including Respondent 1 (DSN, 31), Respondent 2 (PBI, 63), and Respondent 3 (LAS, 63), concerning the presence of light verb constructions. To ensure the accuracy of these findings, an advanced consultation procedure was implemented, namely idiomatic meaning analysis, which involved scrutinizing Goddard & Wierzbicka's (2016) framework.

the genuine dataset, derived through introspective analysis with native Indonesian speakers, embodies the “from within” view. This introspective approach taps into native speakers’ linguistic intuition, capturing potentially novel or less documented LVCs that might be overlooked by relying solely on existing literature. This is crucial for achieving a more comprehensive understanding of Indonesian LVCs, acknowledging the nature of language as perceived by its users. Collectively, the comparison between the hypothetical and genuine datasets, representing these “without” and “within” perspectives, enables a comprehensive investigation into the alignment (or divergence) between theoretical assumptions and empirical observations of Indonesian LVCs.

From a methodological standpoint (Gries & Paquot, 2020; Eatough & Tomkins, 2022), the use of both hypothetical and genuine datasets presents several advantages. Firstly, the hypothetical dataset provides a structured and replicable methodology for identifying potential LVCs. By drawing upon established lists of light verbs from existing literature, the research minimizes subjectivity in the initial identification process. This enhances the reliability of the study by ensuring that the starting point for corpus analysis is transparent and verifiable. Secondly, the genuine dataset offers a complementary approach that allows for the exploration of data patterns. While the hypothetical dataset provides a theory-driven framework, the genuine dataset allows for the discovery of LVCs that might not have been anticipated by prior research. This combination of deductive and inductive approaches strengthens the validity of the findings by providing a more holistic view of LVCs in Indonesian. This mix-method approach in the query assemblage contributes to the robustness of the analysis by mitigating potential biases associated with relying solely on pre-existing knowledge or intuition. Lastly, comparing the results obtained from the hypothetical and genuine datasets allows for a more nuanced interpretation of the corpus data. Similarities between the two datasets can provide converging evidence for the existence and characteristics of certain LVCs, while differences can highlight areas where further investigation is needed.

In addition to these theoretical and methodological considerations, the contextual backdrop of the languages involved plays a crucial role in the dataset assemblage strategy. Specifically, the categorization of languages as either *low-resource* or *rich-resource* significantly influences how linguistic data is obtained and analyzed. Indonesian, in this context, is considered a low-resource language. LRLs are often characterized by a scarcity of digitalized linguistic resources, limited availability of annotated data, and less developed processing tools and technologies. This relative lack of resources poses unique challenges for corpus-based linguistic research. For example, the availability of large, high-quality corpora, which are essential for LVCs analysis, may be limited. Furthermore, the development of reliable NLP tools may be less advanced for LRLs compared to rich-resource languages. Therefore, the reliance on introspective data, as captured in the genuine dataset, becomes particularly important for LRLs like Indonesian, as it provides a valuable source of information that complements the potentially limited corpus data. Conversely, rich-resource languages, such as English, Japanese, or Mandarin, typically boast a wealth or linguistic resources, including large corpora, sophisticated NLP tools, and extensive annotated data. This abundance of resources allows for more data-intensive approaches to linguistic analysis, often relying heavily on statistical modeling and machine learning techniques. In the context of the present study, the contrast between Indonesian as an LRL and other, potentially rich-resource languages from which the hypothetical dataset is derived necessitates a balanced approach. This contextual awareness ensures that the research methodology is appropriate and effective given the specific linguistic landscape.

2.3.1.2 Inclusion and exclusion criteria for dataset assemblage

All datasets underwent an initial screening to validate their forms within the four corpora. This validation process involved checking for the presence or absence of each potential LVC within the four corpora (Baker & Egbert, 2016), with a minimum threshold of one occurrence in any corpus. This seemingly low threshold of a single occurrence is justified by several principles within quantitative and corpus linguistics. Firstly, even infrequent occurrences can be linguistically significant, particularly in large corpora, as they may represent emerging trends or less common but still valid constructions (Baker *et al.*, 2006). Secondly, in corpus linguistics, the focus is not solely on the most frequent items, but also on the range and distribution of linguistic phenomena (McEnery & Hardie, 2012). Therefore, including items with at least one occurrence ensures that the analysis captures a broader spectrum of potential LVCs. After this validation, the genuine dataset, initially comprised of 102 items, was refined to 101 items. Similarly, the hypothetical dataset, which initially contained 900 items, was reduced to 841 items following the validation procedure. This screening process yielded the validated datasets of the two matrices, ready to be used in the further analysis (*see* sample list in Table 2.3).[12]

[12] This version constitutes a dataset that meets the fundamental principle known as “the golden rule of data.” As Wallis (2020: 263) states, “Whether one is using a simple lexical list query or the most complex grammatical

Table 2.3: Sample list of the hypothetical and genuine matrices.

Hypothetical Matrix	Genuine Matrix
memberi arahan 'gives direction'	*masuk akal* 'to make sense'
membuat salinan 'make a copy'	*campur tangan* 'to interfere'
mengambil bagian 'take part'	*pasang surut* 'ups and down'
melakukan perekaman 'get video'	*bertolak belakang* 'to be contrary'
membuat lelucon 'make a joke'	*gulung tikar* 'to out of business'
memberikan ulasan 'make comment'	*turun tangan* 'to take action'
mengindahkan perhitungan 'takes into account'	*putus sekolah* 'to drop out of school'
mengambil catatan 'take notes'	*pandang bulu* 'to show favoritism'
membuat keputusan 'make a decision'	*menutup mata* 'passing away'
membawa uang 'brings money'	*menjauhkan diri* 'to distance oneself'
membuat penawaran 'make an offer'	*terbawa arus* 'to follow the crowd'
melakukan pendaftaran 'do registration'	*gigit jari* 'to bite one's fingers'
memberi kesempatan 'give an opportunity'	*naik daun* 'to rise in popularity'
memberi contoh 'gives an example'	*tertangkap basah* 'to be caught red-handed'
memberikan izin 'give a permission'	*membuka jalan* 'to pave the way'
memberikan nasihat 'give advice'	*patah semangat* 'to lose spirit'
melakukan lawatan 'make a visit'	*cuci darah* 'to cleanse one's lineage'
memberikan kepuasan 'give a satisfaction'	*mengadu nasib* 'to try one's luck'
memberikan pujian 'give compliment'	*menarik diri* 'to pull out'
membuat kesepakatan 'conclude an agreement'	*jatuh hati* 'to fall in love'
memberi arahan 'gives direction'	*memutar otak* 'thinking out loud'
mengambil bagian 'take part'	*berpangku tangan* 'to sit idly by'
mengadakan kontrak 'enters into a contract'	*adu domba* 'to pit against each other'
mengambil jarak 'take distance'	*mengulurkan tangan* 'to lend a hand'
memberikan paparan 'give a presentation'	*mati rasa* 'to feel nothing'

The data validation also conducted based on inclusion and exclusion criteria for data assemblage. These criteria or parameters were as follows: (a) semantic contribution of the noun: to align with Mel'čuk's (2022) emphasis on the noun carrying primary meaning, constructions were included only if the noun constituent contributed the core semantic predicate of the phrase; LVCs where the verb appeared to carry significant semantic weight, overshadowing the noun's contribution to the overall meaning, were excluded; (b) grammatical placeholder function of the verb: following Mel'čuk's (2022) definition of the light verb as "grammatical placeholder," constructions were included if the verb functioned primarily to support the expression of

query, we must always hold to the same fundamental principle: The golden rule of data. We need to know that, as far as possible, our dataset is a sound and complete set of examples of the linguistic phenomenon in which we are interested."

grammatical functions, rather than contributing substantial independent semantic content; verb-noun combinations where the verb retained a strong independent meaning were excluded; and (c) paraphrasability test: to further ensure adherence to Mel'čuk's (2022) characterization, a paraphrasability test was applied; constructions were included if the meaning of the verb-noun phrase (VN) was essentially the same as that of a full verb derivationally related to the noun (V_o); if such a paraphrase was not possible, the construction was excluded.

2.3.1.3 Frequency extraction procedures

The subsequent step of data collection is frequency extraction which consists of the following three steps. First, initial retrieval and close reading. The initial phase of frequency extraction involved a meticulous and systematic retrieval of each construction listed in the two datasets (hypothetical and genuine) across the four corpora (ILCC, SLIC, IDC, and IWC). This process was not automated; rather, it necessitated a manual approach to ensure accuracy and account for the nuances of linguistic variation (*see* Figure 2.3). The methodology was heavily informed by Mehl's (2019) corpus semantic approach, specifically the concept of "identity evidence." Mehl's work underscores the importance of close reading in corpus linguistics, advocating for a nuanced examination of contextual cues to accurately identify lexical items or unites within a corpus. This is particularly crucial when dealing with constructions like LVCs, where the light verb and the noun may appear in various forms and grammatical contexts. "Identity evidence," as Mehl (2019) explains, involves a careful consideration of the surrounding words, grammatical structures, and semantic context to confirm that a particular instance within the corpus truly represents the LVC under investigation. This close reading approach was crucial in mitigating the challenges posed by polysemy (words with multiple meanings) and homonymy (words with the same form but different meanings), ensuring that only relevant instances of the LVCs were counted. The manual retrieval, guided by this "identity evidence" principle, allowed for a more precise and reliable extraction of LVC occurrences from the corpora.

Second, compilation of LVCs frequency. Following the initial retrieval, the subsequent step involved the compilation of a comprehensive list of the LVCs' frequency data. This process aimed to quantify the extent of their occurrence within each corpus. The identified LVCs, validated and extracted through the close reading method, were carefully tabulated to create a structured record of their frequency. This tabulation was designed to capture frequency across all corpora, providing a rich dataset for subsequent analysis. Each LVC entry in the table included the total frequency, representing the raw count of how many times that specific LVC appeared in the corpus. This raw count provides a basic measure of the LVC's prevalence. In addition to the total frequency, the data was also converted into percentages, allowing for a comparison of the relative frequency of different LVCs within the same corpus. Furthermore, a rank was assigned to each LVC based on its frequency, providing a clear indication of its relative prominence compared to other LVCs in that corpus.

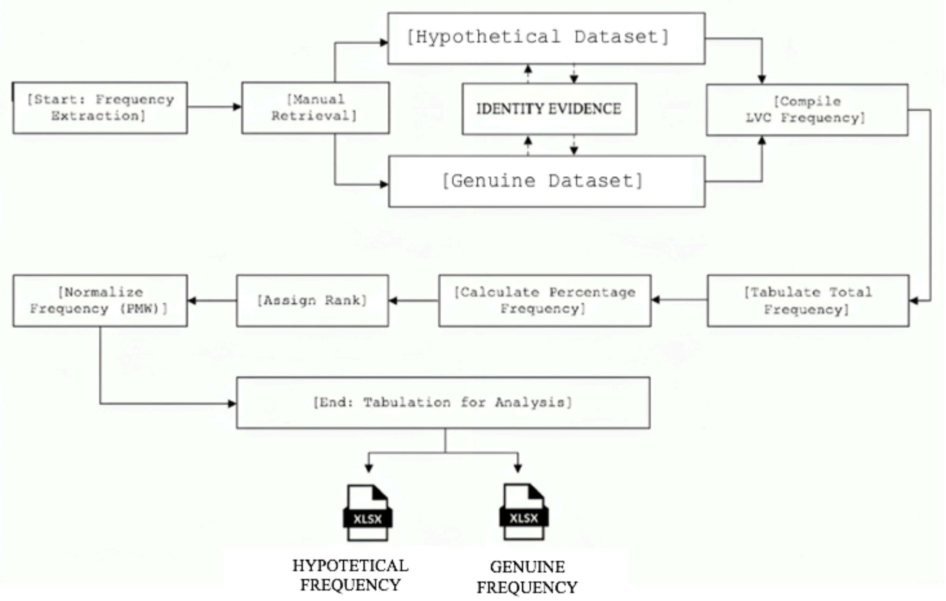


Figure 2.3: Frequency extraction procedures.

Lastly, normalization and tabulation for analysis. Normalization is a crucial step in corpus linguistics, as it allows researchers to compare frequency data across corpora of different sizes (Baroni & Evert, 2009; Biber & Jones, 2009). The PMW measure provides a standardized metric that allows for a more accurate comparison of LVC frequencies across the ILCC, SLIC, IDW, and IWC, despite their varying sizes. The comprehensive tabulation, encompassing total frequency, percentage, rank, and PMW, served as a foundational resource for further quantitative analysis. This meticulously prepared dataset enabled a range of statistical analyses, including comparisons of LVC frequencies across different corpora, identification of the most frequent LVCs, and analysis of the distribution patterns of LVCs in a subsequent grammatical-context. The tabulated data provided a solid empirical basis for investigating the research questions, allowing for an empirical exploration of the characteristics and behavior of LVCs in Indonesian.

2.3.2 Data analysis procedures

2.3.2.1 Frequency analysis

Frequency data provides a crucial quantitative measure of the prevalence and distribution of linguistic feature within a corpus (McEnery & Brezina, 2022). The frequency analysis of LVCs was conducted following the principles inspired by the procedure outline by McEnery and Hardie (2012). The analysis involved generating frequency lists for LVCs identified in both the hypothetical and genuine datasets across the four corpora. This process yielded raw frequency counts, indicating the absolute number of occurrences of each LVC. However, to facilitate meaningful comparisons between corpora of varying sizes, the raw frequencies were normalized. As Baker *et al.* (2006: 75) further explain, frequencies can be presented as percentages or proportions, or more practically, as occurrences per-million-words (hereafter: PMW), allowing for standardized comparison across corpora of different dimensions. This normalization process is essential for mitigating the skewing effects of corpus size on raw frequency counts.

To further analyze the circulation of LVCs within the all datasets, a process of frequency clusterization was employed by utilizing K-means Clustering performed in R-Studio application.[13] K-means clustering, a prominent unsupervised machine learning algorithm, functions by partitioning a dataset into *k* distinct, non-overlapping subgroups, or clusters. This partitioning is achieved iteratively, wherein data points are assigned to the cluster with the nearest centroid, followed by a recalculation of the centroid as the mean of all points within that cluster. The algorithm converges upon a solution that minimizes the within-cluster sum of squared errors, effectively maximizing intra-cluster similarity and inter-cluster dissimilarity. The adoption of *k*-means clustering for the analysis of LVCs is predicated upon several key advantages. Firstly, its computational efficiency renders it scalable to the substantial datasets frequently encountered in corpus linguistic research. Secondly, *k*-means requires no a priori knowledge of class labels, allowing for the data-driven emergence of usage patterns. Thirdly,

[13] R-Studio (version 2024.2.1) has been used to statistically calculate the homogeneity of frequency data to determine the cluster of Indonesian LVCs based on K-Means Clustering. The employment of R-Studio facilitates the implementation of *k*-means clustering through its robust statistical computing environment and extensive suite of packages designed for data analysis, thereby ensuring methodological rigor in the identification of homogeneous LVC groupings. Furthermore, the capacity of R-Studio to handle large datasets and perform complex calculations, including the iterative centroid adjustments inherent in *k*-means clustering, renders it particularly suitable for processing the substantial corpus data required to derive reliable clusters of Indonesian LVCs. Finally, the visualization capabilities of R-Studio, enabling the graphical representation of cluster assignments and homogeneity metrics, contribute to a more transparent and interpretable presentation of the resultant LVC groupings, enhancing the overall clarity of the analysis.

the method's centroid-based representation offers an intuitive means of characterizing cluster profile, facilitating linguistic interpretation of the identified groupings.

Accordingly, the analysis involved stratifying the LVCs into three distinct categories based on their homogeneity of frequency: (a) high-frequency, (b) medium-frequency, and (c) low-frequency. The high-frequency category, characterized by LVCs exhibiting the highest frequency across the corpora, denotes their relatively frequent occurrence within the four corpora under examination, and is assigned code of Q3. Conversely, the low-frequency category coded as Q1, encompasses LVCs with lower frequency, indicative of their infrequent occurrence. The medium-frequency category, coded as Q2, refers to LVCs demonstrating frequency within the intermediate range. The specific thresholds for these categories were determined empirically, calculated algorithmically based on the K-means clustering of the frequency LVCs within all dataset across the collective corpus (Gries, 2015a, 2015b; Sheng, 2023).

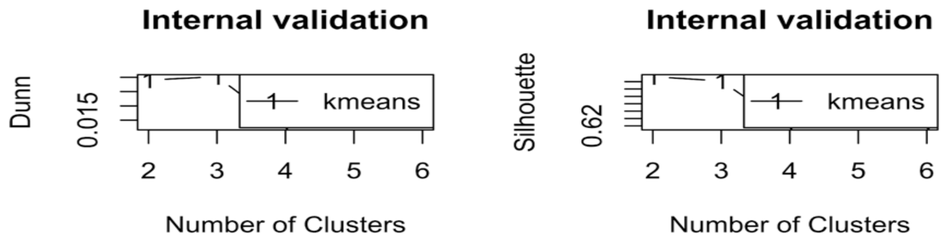


Figure 2.4: Dunn-index evaluation for K-means clustering of LVCs.

Based on the validation measures, there are several important justifications in determining the three clusters. First, as illustrated in Figure 2.4, Dunn evaluation for optimal number of clusters. To validate the efficacy of the K-means clustering procedure in partitioning the LVCs into three distinct groups, the Dunn Index, a metric that evaluates clustering performance by considering both inter-cluster separation and intra-cluster compactness, was employed. Specifically, the Dunn Index quantifies the ration of the minimal inter-cluster distance to the maximal intra-cluster distance; a higher Dunn Index value signifies superior clustering, indicative of well-separated and internally cohesive groups. While the Dunn Index provides a global assessment of cluster quality, supplementing this analysis with Silhouette scores offers more granular perspective. Silhouette analysis assesses the degree to which each individual data point is appropriately assigned to its cluster. Silhouette scores, ranging from -1 to 1, reflect the difference between a data point's average distance to other points in its own cluster and its average distance to points in the nearest neighboring cluster. Scores approaching

1 suggest that point is well-clustered, whereas scores near-1 imply misclassification. Therefore, the combined interpretation of a favorable Dunn Index, coupled with a distribution of Silhouette scores demonstrating a preponderance of values above zero, provides robust evidence for the appropriateness of the K-means clustering into specified three groups.

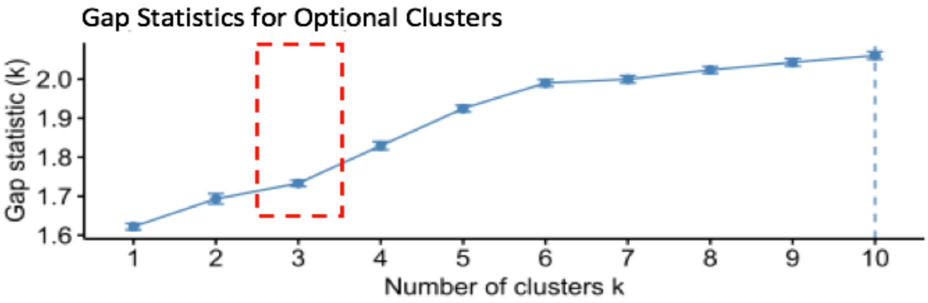


Figure 2.5: Gap-statistics evaluation for K-means clustering of LVCs.

Second, Gap Statistics for optimal clusters as illustrated in Figure 2.5. The determination of the optimal number of clusters for the K-means partitioning was also facilitated by the Gap Statistics, a methodology that compares the within-cluster dispersion of the observed data to that expected from a uniform distribution. Specifically, the Gap Statistics computes the difference between the logarithm of the within-cluster dispersion for the actual data and the average of the logarithm of the within-cluster dispersion for a set of reference datasets generated under a null hypothesis of no discernible clustering. The optimal number of clusters is identified as the smallest *k* for which the Gap Statistics is sufficiently larger than the Gap Statistics for *k* + 1. As demonstrated in Figure 2.5, a primary discernible "preferably gap" is observed at *k* = 3, indicating that three clusters represent a statistically justifiable configuration. This approach mitigates the inherent subjectivity of selecting *k,* providing a statistically grounded criterion for ascertaining the most appropriate cluster configuration.

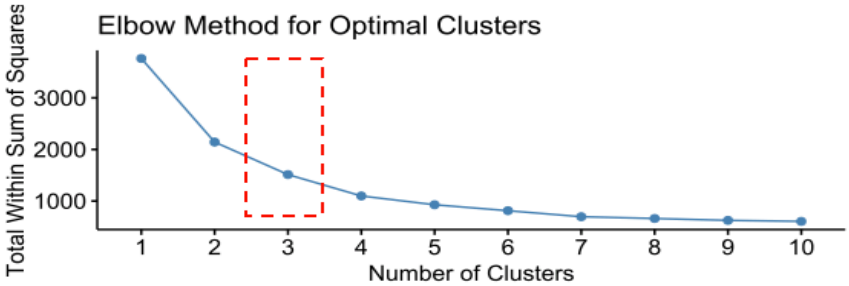


Figure 2.6: Elbow-method evaluation for K-means clustering of LVCs.

Lastly, Elbow method for optimal clusters. This method was also employed to ascertain the optimal number of clusters for the K-means partitioning of LVCs. This technique involves plotting the within-cluster sum of squared errors as a function (WCSS) of a function of *k*. Initially, WCSS decreases sharply with increasing *k*, as each additional and their respective centroids. However, beyond a certain *k,* the rate of decrease in WCSS diminishes significantly, forming an 'elbow' in the plot. The *k* value corresponding to this elbow is conventionally interpreted as the point of diminishing returns, indicating that further increases in *k* yield only marginal reductions in within-cluster variance. As depicted in Figure 2.6, the Elbow method analysis suggests $k = 3$ as the optimal number of clusters, signifying that three groups effectively capture the inherent structure of the LVC data without introducing excessive fragmentation.

2.3.2.2 Distribution analysis

To investigate the distribution of Indonesian LVCs, it was necessary to consider the previous study conducted by Nugraha (2024) and Nugraha and Vincze (2024) on specific LVC. These studies were examined the distribution of canonical type of LVCs, namely 'give,' 'take,' and 'have*,*' as stated by Jespersen (1965: 117). The findings of the preliminary study suggest the existence of Indonesian LVCs outside the canonical category. Consequently, further exploration into the full range of Indonesian LVCs is warranted. To facilitate the exploration, a practical methodological step involves the development of a data analysis instrument. Based on the aforementioned considerations, instruments were developed to support the investigation of Indonesian LVCs distribution across six predefined conditions or contexts.

The first context was measured using the M-1 parameter, which focused on tracing the distribution of LVCs based on the nominal morphological features as embedded in the noun element (Mel'čuk, 2006; Blevins, 2016; Lieber & Štekauer, 2011). The second context was measured using the M-2 parameter, which focused on tracing the distribution of LVCs based on their primary feature of the noun element, namely the conceptual skeleton of SUBSTANCE (Lieber, 2004, 2010; Dal & Namer, 2015). The third context was measured using the S-1 parameter, which focused on tracing the distribution of LVCs based on their scale of synonymity (Cruse, 1986; Andreou, 2017; Hrenek, 2021). The fourth context was measured using the S-2 parameter, which focused on tracing the distribution of LVCs on the prototypicality (Coleman & Kay, 1981; Singleton, 2000; Aikhenvald, 2006, 2007, 2017). The last two were Sx-1 for the examination of LVC within the transitivity parameter (Hopper & Thompson, 1980; Malchukov, 2006; Kittilä, 2006) and Sx-2 for the examination of LVC within the valency frame (Kettnerová, 2023; Vincze, 2014). These six types of analysis were used to describe the distribution of LVCs in Indonesian grammatical contexts.

The following provides a detailed description of each parameter. Initially, the M-1 parameter is designed to delineate the distribution of LVCs based on the nominal morphological features inherent in the noun element. The concept of nominal morphological features refers to Mel'čuk's (2006) formulation of the linguistic sign, outlined below:

> "A *linguistic sign* is an ordered triplet
>
> $$X = <^{(}X^{)} ; /X/ ; \Sigma_{X}>,$$
>
> where $^{(}X^{)}$ is a signified, /X/ a corresponding signifier, and Σ_{X} is the syntactics of the pair $<^{(}X^{)} ; /X/>$.
>
> Here, signified (= signatum, *signifié*) and signifier (= signans, *signifiant*) are taken in their Saussurian sense; syntactics is a set of all combinatorial properties, or features, of the sign X that are determined neither by its signified nor by its signifier. These are such features as part of speech, inflectional class, agreement class (in particular, grammatical gender or noun class), government pattern, possible syntactic constructions, restricted lexical co-occurrence, etc. Linguistic signs include: simple morph, derived stems, wordforms, free phrases, set phrases (phrasemes), reduplications, apophonies, conversions."
>
> (Mel'čuk, 2006: 384-385)

The M-1 parameter, designed to delineate the distribution of LVCs based on the morphological signs inherent in the noun element, draws heavily from Mel'čuk (2006) formulation of the linguistic sign. Mel'čuk conceptualizes a linguistic sign as an "ordered triplet," comprising three key components: the signified ((X)), the signifier (/X/), and the syntactics of the pair (Σx). The signified refers to the concept or meaning that the sign represents. The signifier is the form that the sign takes, such as a sound sequence or a written word. Crucially, Mel'čuk introduces syntactics (Σx) as a set of combinatorial properties or features of the sign that are independent of both its signified and its signifier. These syntactic features encompass a range of grammatical and structural characteristics. Within the context of the M-1 parameter, this conception of the linguistic sign is applied to analyze the distribution of LVCs based on the morphological signs of their noun components. To achieve this, two primary categories are delineated: LVCs with 'Base' (Ba) nouns, exhibiting no morphological affixation, and those with 'Affixed' (Af) nouns.

This categorization allows for a systematic comparison of how the presence or absence of morphological affixes on the noun influences the behavior and distribution of LVCs.

To analyze LVCs distribution within the M-1 parameter using manual close reading, a structured technique was implemented. First, each instance of LVCs from both the hypothetical and genuine datasets was prepared. Second, the noun element within each LVC instance was identified. Third, a meticulous examination of the morphological structure of each noun element was conducted. This involved identifying and categorizing any morphological affixes present, such as prefixes, suffixes, infixes, or circumfixes. Nouns lacking any such affixes were classified as 'Base' (Ba) nouns. Fourth, each LVC instance was categorized based on whether its noun element was classified based on whether its noun element was classified as 'Base' (Ba) or 'Affixed' (Af). Fifth, the distribution of LVCs across these two categories was tabulated separately for the hypothetical and genuine datasets, noting the frequency of LVCs with 'Base' nouns and those with 'Affixed' nouns. Finally, a comparative analysis of these distributions was performed to discern any significant differences in the occurrence of 'Base' and 'Affixed' nouns within LVCs across the two datasets.

Second, the M-2 parameter is designed to trace the distribution of LVCs based on their primary feature of the noun element, namely conceptual skeleton of SUBSTANCE (Lieber, 2004, 2010; Dal & Namer, 2015). Drawing upon descriptive framework developed by Lieber (2004), a salient feature of the noun element that can be assigned as the basis for M-2 parameter is the most basic conceptual skeleton of SUBSTANCES/THINGS/ESSENCES. Specifically, the skeleton characterized by [+/- material], as Lieber (2004) explains:

> "[+/- material]: The presence of this feature defines the conceptual category of SUBSTANCES/THINGS/ESSENCES, the notional correspondent of the syntactic category of Noun. The positive value denotes the presence of materiality, characterizing concrete nouns. Correspondingly, the negative value denotes the absence of materiality; it defines abstract nouns."
>
> (Lieber, 2004: 24)

Drawing from Lieber's (2004) descriptive system, the core concept employed here is the conceptual skeleton of SUBSTANCES/THINGS/ESSENCES, a category that aligns with the syntactic category of Noun. Crucially, this conceptual skeleton is characterized by the binary value [+/- material], a distinction that allows for the classification of nouns based on their inherent materiality. The presence of the feature [+material] signifies materiality, thereby defining concrete nouns. Concrete nouns (Co), in this context, refer to tangible objects or entities that possess physical substance and can be perceived through sensory experience.

Examples of concrete nouns in Indonesian include '*buku*' (book) and '*meja*' (table), which denote physical, touchable entities. Conversely, the absence of materiality, indicated by the feature [-material], defines abstract nouns (Ab). Abstract nouns represent intangible concepts, ideas, or qualities that lack physical substance and cannot be directly perceived by the senses. In Indonesian, examples of abstract nouns include '*kebahagiaan*' (happiness) and '*kebebasan*' (freedom), representing non-physical, conceptual entities. This distinction between concrete and abstract nouns, underpinned by the [+/- material] feature, is central to the M-2 parameter.

To analyze LVCs distribution within the M-2 parameter using manual close reading, a detailed stage was implemented. First, all instances of LVCs were prepared from both the hypothetical and genuine datasets. Second, for each LVC instance, the constituent noun element was identified. Third, the noun element was analyzed to determine its primary morphological-feature, focusing on the concept of materiality. This involved assessing whether the noun denoted a tangible entity with physical substance (concrete, [+material]) or an intangible concept lacking physical substance (abstract, [-material]). Fourth, each LVC instance was classified based on the materiality of its noun element, being categorized as either 'Concrete' (Co) or 'Abstract' (Ab). Fifth, the distribution of LVCs across these two categories was tabulated separately for the hypothetical and genuine datasets. This tabulation involved recording the frequency with which LVCs occurred with concrete nouns and abstract nouns in each dataset. Finally, a comparative analysis of these distributions was conducted to identify any significant differences in the patterns of LVC usage with concrete versus abstract nouns between the hypothetical and genuine datasets.

Third, the S-1 parameter is designed to trace the distribution of LVCs according to their *scale of synonymity* (Cruse; 1986). Cruse (1986) explains this concept as follows:

> "Within the class of synonyms, some pairs of items are more synonymous than others, and this raises the possibility of a scale of synonymity of some kind. A scale needs at least one well-defined end-point; and if there is only one, it is more satisfactory for it to form the origin, or zero point, on the scale. With regard to degrees of synonymity, it seems that the point of semantic identity – i.e. absolute synonymy – can be established with some clarity; the notion of zero synonymity, on the other hand, is rather more diffuse. For one thing, it is probably not a unitary concept: *long*:*short* and *green*:*expensive* would probably both count as examples of zero synonymity."
>
> (Cruse, 1986: 267-268)

According to Cruse (1986), synonymy exists on a scale, implying that some construction pairs exhibit a stronger degree of semantic similarity than others. This gradation of synonymy is central to the S-1 parameter. Cruse (1986) argues for the existence of a “scale of synonymity,” a concept that acknowledges the nuanced relationships between words or word-constructions. He highlights the challenge of defining “zero synonymity,” demonstrating that the absence of synonymy is not a uniform concept. While it is relatively easy to understand what it means for words to be very similar (synonymous), it is much harder to pin down what it means for words to be completely different (not synonymous at all). The S-1 parameter draws on Cruse’s concept to analyze the distribution of LVCs by examining their relationships with potential counterpart expressions. The purpose is to calculate the distribution of LVCs based on whether they have a counterpart (Cp) or lack a counterpart (Ncp) within this scape of synonymity. In essence, the S-1 parameter seeks to determine the degree to which an LVC has semantically similar alternatives. By analyzing LVCs along Cruse’s scale of synonymity, this parameter helps to reveal the semantic flexibility or rigidity of these constructions. It allows us to investigate whether LVCs tend to cluster at specific points on the synonymy scale, indicating a preference for certain types of semantic relationships.

To technically analyze the distribution of LVCs within the S-1 parameter using manual close reading, a set of steps was employed. First, each instance of the LVCs from both the hypothetical and genuine dataset was prepared. Second, for each extracted LVC, a comprehensive search was conducted to identify potential alternative expressions, or “counterparts” (Cp), that could convey similar meaning. This involved examining dictionaries, thesauruses, and linguistic resources, as well as considering contextual information from the native Indonesian. Third, the semantic relationship between each LVC and its identified counterpart(s) was carefully analyzed. Fourth, LVCs were categorized based on whether they had a clear counterpart (Cp) or lacked a clear counterpart (Ncp) within the given context. If multiple counterparts were identified, the LVC was assessed based on the closest synonymic relationship. Fifth, the distribution of LVCs across the categories (Cp or Ncp) was tabulated for both the hypothetical and genuine datasets. Finally, the distributional patterns were analyzed to determine the prevalence of LVCs with and without counterparts, and to identify any significant differences between the datasets.

Fourth, the S-2 parameter is informed by the concept of *prototypicality* (Coleman & Kay, 1981; Singleton, 2016; Taylor, 2015), designed to trace the distributional patterns of LVCs in relation to their core prototypical meaning, as determined by their constituent noun element. Coleman and Kay (1981) elucidate this concept as follows:

> "Our *prototype* view of word meaning attempts to account for obvious pretheoretical intuition that that semantic categories frequently have blurry edges and allow degrees of membership. On this view, applicability of a word to a thing is general not a matter of 'yes or no', but rather of 'more or less'. A semantic prototype associates a word or phrase with a prelinguistic, cognitive schema or image; that speakers are equipped with an ability to judge the degree to which an object matches this prototype schema or image."
>
> (Coleman & Kay, 1981: 28)

This conceptualization provides a crucial theoretical foundation for the S-2 parameter. The concept of prototypicality addresses the inherent variability and graded nature of word meaning, acknowledging that semantic categories are not always clearly defined and that membership within a category is often a matter of degree rather than absolute inclusion or exclusion. This perspective is particularly relevant for analyzing LVCs, where the semantic contribution of the noun component can significantly influence the overall meaning and usage of the construction. The S-2 parameter leverages this notion of prototypicality by classifying nouns based on their prototypical meanings. In this analysis, nouns can be characterized into several classes, for instance *people* (Pe), *plants* (Pl), *animal* (An), *material* (Ma), *objects* (Ob), *qualities* (Qu), *action* (Ac), and *processes* (Pr) (*see* Aikhenvald, 2006, 2007, 2017). This classification aims to capture the core semantic essence of nouns, reflecting how they are typically conceptualized and categorized by language users. By examining the distributional patterns of LVCs across these semantic categories, this part of analysis seeks to identify the semantic constraints and possibilities that govern their usage. The significance of prototypicality for LVC analysis lies in its ability to account for the semantic flexibility and variability observed in these constructions. For instance, the same light verb might exhibit different distributional patterns depending on whether it combines with a noun denoting an *action* or a noun denoting a *quality.* This variation reflects the influence of the noun's prototypical meaning on the overall semantics of the LVC.

Furthermore, to analyze the LVC distribution within the S-2 parameter using manual close reading, systematic steps were employed. First, each instance of the LVCs from both hypothetical and genuine datasets was prepared. Second, the noun component of each LVC instance was identified. Third, each identified noun was carefully analyzed to determine its core prototypical meaning. This involved considering the typical usage and conceptualization of the noun. Fourth, based on this analysis, each noun was categorized into one the predefined semantic classes (People, Plants, Animal, Material, Objects, Qualities, Action, or Processes). Fifth, the distribution of LVCs across these semantic categories was tabulated, noting how frequently each LVC occurred with nouns from each category. Finally, the distributional patterns were examined, comparing the hypothetical and genuine datasets to identify any significant differences in how LVCs combine with nouns from different semantic classes.

Fifth, the Sx-1 parameter is designed to delineate the distributional patterns of Indonesian LVCs within the spectrum of transitivity. This parameter is derived from Hopper and Thompson's (1980) influential work on transitivity parameters. In their notable contribution to linguistic theory, Hopper and Thompson (1980) conceptualize transitivity as follows:

> "Transitivity, viewed in the most conventional and traditional way possible—as a matter of carrying-over or transferring an action from one participant to another—can be broken down into its component parts, each focusing on a different facet of this carrying-over in a different part of the clause. Taken together, they allow clauses to be characterized as *more* or *less* Transitive: the more features a clause has in the 'high', the more Transitive it is—the closet it is to cardinal Transitivity."
> (Hopper & Thompson, 1980: 253)

Hopper & Thompson (1980) offer an important perspective on transitivity, moving beyond the simplistic notion of an action merely being transferred from one participant to another (*see also* Lieber, 2006; Roßdeutscher, 2014). As the basis for the Sx-1 parameter, their framework provides a valuable tool for analyzing LVCs. They propose that transitivity is not a binary concept but rather a continuum, determined by a cluster of interrelated parameters (*see* Table 2.4). These parameters collectively characterize a clause as being "more or less Transitive." Each parameter focuses on a distinct facet of the "carrying-over" of an action within different parts of the clause. High transitivity is associated with features such as having two participants, action verbs (kinesis), telic aspect (having a defined endpoint), punctual events, volitional actors, affirmative clauses, realis mode (factual), high agency of the subject, a totally affected object, and a highly individuated object. Conversely, low transitivity involves the opposite features. By breaking down transitivity into these components, Hopper and Thompson (1980) provide a more granular way to assess the degree to which an event is effectively transferred from an agent to a patient. In the context of LVC analysis, this framework allows for a detailed examination of how light verbs and their associated nouns interact to influence the overall transitivity of the construction.

Table 2.4: Hopper and Thompson's (1980) Transitivity Parameters.

Hopper & Thompson's (1980) Transitivity Parameters		
Parameter	*High*	*Low*
A. Participants	2	1
B. Kinesis	Action	Non-action
C. Aspect	Telic	Atelic
D. Punctuality	Punctual	Non-punctual
E. Volitionality	Volitional	Non-volitional
F. Affirmation	Affirmative	Negative
G. Mode	Realis	Irrealis
H. Agency	A high in potency	A low in potency
I. Affectedness of O	O totally affected	O not (totally) affected
J. O individuation	O highly individuated	O non-individuated

To analyze the distribution of LVCs within the Sx-1 parameter using manual close reading, specific steps were employed. First, each instance of the LVCs from both the hypothetical and genuine datasets was prepared. Second, for each instance, the sample clause containing the LVC was examined to identify the presence and nature participants, the type of action conveyed (kinesis), the aspectual properties (telic or atelic), punctuality, volitionality of the agent, affirmation or negation, mode, agency of the subject, affectedness of the object, and individuation of the object. Third, each of these parameters was assessed and categorized based on the "high" and "low" transitivity features as defined by Hopper and Thompson (1980). Fourth, the overall transitivity of each LVC instance was determined by considering the collective configuration of these parameters. Finally, the distributional patterns of LVCs across the transitivity continuum were analyzed, comparing the hypothetical and genuine datasets to identify variations in their transitivity profiles.

As the last, the Sx-2 parameter, termed *valency scope* in this study and closely related to Kettnerová's (2023) concept of *valency frame,* is a crucial parameter for analyzing the distributional properties of Indonesian LVCs. In detail, Kettnerová (2023) explains as follows:

> "The valency structure of a single word predicate (typically a verb, a noun or an adjective) is captured by the *valency frame* stored in a lexicon. The valency frame consists of a set of valency slots, each filled with one valency complementation. Only actants (be they obligatory or optional) and free modifications that are obligatory are part of the valency frame (optional free modifications stand outside the valency frame). Each valency complementation is assigned with a functor and with information on obligatoriness."
>
> (Kettnerová, 2023: 23)

Valency, in essence, captures the capacity of a predicate, typically a verb in LVCs context, to govern other elements within a clause or sentence. This governing capacity is formally represented by a *valency frame,* which is a structured set of *valency slots.* Each slot is filled by a *valency complementation,* representing the linguistic elements that are required or allowed by the predicate. These valency complementations can include actants, which are the core participants in the event or state described by the predicate, and obligatory free modifications, which, while not actants, are nonetheless required for the completeness of the predication. Crucially, optional free modifications are excluded from the valency frame, as they are considered to lie outside the predicate's inherent governing requirements. Each valency complementation within the frame is further specified by functor, denoting its semantic role (e.g., agent, patient, instrument), and information pertaining to its obligatoriness, indicating whether the complementation is required or optional for a grammatically complete and semantically coherent predication. Therefore, the Sx-2 parameter, by focusing on valency scope, facilitates a detailed analysis of the syntactic behavior or Indonesian LVCs, revealing the range of complements they govern.

To capture this complexity, the current investigation examines the current type of analysis examines the distributional patterns of LVCs within two distinct valency paradigms: those exhibiting a fixed valency, semantically tending toward idiomaticity, and those demonstrating a more flexible valency, semantically exhibiting greater compositionality. LVCs with a fixed valency often mirror the behavior of their constituent light verbs, displaying a relatively rigid set of obligatory and optional complements, akin to idiomatic expressions where the meaning is less predictable from the individual parts. In contrast, LVCs with a flexible valency are more influenced by the semantic and syntactic properties of the incorporated noun, allowing for a wider range of possible complements and exhibiting greater compositionality, where the meaning of the phrase is more readily derived from its constituent elements. This distinction allows for a more specific analysis of how LVCs function within the syntactic structure of Indonesian.

To analyze the distribution of LVCs within the Sx-2 parameter using manual close reading, the following steps were undertaken. First, each instance of the LVCs from both the hypothetical and genuine datasets was extracted from the corpora. Second, the immediate syntactic context surrounding each LVC occurrence was carefully examined to identify the elements governed by the LVC. Third, the valency frame for each LVC instance was constructed, noting the actants and obligatory free modifications present, either as fixed or flexible valency. Fourth, the semantic roles of the identified valency complementations were determined, assigning appropriate functor (e.g., agent, patient, instruments). Finally, the obligatoriness of each complementation was assessed, noting whether its presence was required for grammaticality and semantic completeness of the clause.

2.3.2.3 Verb and noun elements analyses

The subsequent part of analysis is to examine the aktionsart of verb elements within Indonesian LVCs. The specific parameter applied in this stage, i.e. metric for aktionsart detection that only can be done by human operations. As such, an intuition of the author as native Indonesian is played significantly. The metrices have been adopted from (a) Binnick's (1991) work for verb element and (b) Pompei *et al.*'s (2023) study for noun element. Table 2.5 and 2.6 presents the matric for this purpose of analysis. Primarily, the analysis of the verb element represents a crucial aspect of this investigation, necessitating a detailed examination that differentiates between True Light Verbs (hereafter: TLVs) and Vague Action Verbs (hereafter: VAVs). This distinction, as informed by the metric outlined in Table 2.5, is not merely a categorial exercise, but rather a fundamental step towards unravelling the intricate morphosyntactic and semantic properties that govern LVC formation and interpretation. This framework employed herein leverages a set of inherent temporal properties—namely: momentary, durative, frequentative, iterative, and intensive—to systematically evaluate the contribution of the verb element to the overall aspectual profile of the LVC.

Table 2.5: Metric for evaluating the inherent temporal properties of a situation within verb of Indonesian LVCs.

Form of verb	Momentary	Durative	Frequentative	Iterative	Intensive
….	[+/-momentary]	[+/-durative]	[+/-frequentative]	[+/-iterative]	[+/-intensive]
….	[+/-momentary]	[+/-durative]	[+/-frequentative]	[+/-iterative]	[+/-intensive]

Note:
+ : presence
- : absence

TLVs, by their nature, exhibit a diminished semantic content. They often lack or possesses only weakly specified values for the temporal properties under consideration. For instance, a TLV such as '*melakukan*' (to do) typically contributes little to the momentariness, durativity, frequentativeness, iterativeness, or intensiveness of the construction. Instead, their primary function is to provide grammatical support to the noun, enabling it to function as the predicate of the clause. In terms of the metric in Table 2.5, TLVs would ideally show a greater propensity for [-] values across the features, indicating the absence of a string specification for that particular property. This absence of strong temporal specification is a defining characteristic of TLVs, setting them apart from their more semantically robust counterparts. The analysis, therefore, involves a rigorous examination of corpus data to identify instances where the verb exhibits this kind of 'bleached' semantics.

On the other hand, VAVs occupy a somewhat intermediate position. While they may not possess the full semantic weight of main verbs, they do contribute a more discernible degree of meaning to the LVC than TLVs. VAVs typically denote a general type of action or activity, and as such, they often carry some inherent aspectual information. For example, a VAV might specify durativity (e.g., suggesting that the event unfolds over a period of time) or dynamicity (which can be related to momentariness or iterativeness). However, crucially, this aspectual contribution is often less precise and less constrained than that of a full verb, and it's heavily influenced and modulated by the noun within the LVC. Referring to Table 2.5, VAVs might show a mix of [+] and [-] values, reflecting their partial specification for certain temporal properties. For example, a VAV might be [+durative] but [-intensive].

The analysis of the verb element, therefore, proceeds along two interconnected paths. First, it involves a detailed examination of the distributional patterns of verbs within LVCs datasets, focusing on the range of nouns with which they co-occur. TLVs are expected to exhibit a broader range of co-occurrence, combining with a wider variety of noun types, precisely because they contribute less semantic content of their own and are more grammatically driven. Second, the analysis involves a semantic investigation, using the framework in Table 2.5, to assess the degree to which the verb contributes to the momentariness, durativity, frequentatives, iterativeness, and intensiveness of the LVC. This involves careful consideration to the context in which the LVC occurs, to determine whether the temporal properties of the construction are primarily determined by the noun, or whether the verb makes a more substantial contribution. The final stage is to create a categorical tabulation based on the analysis of the verb elements.

Furthermore, the analysis of the noun element constitutes a critical facet of this research, demanding a specific approach that differentiates between Stative and Eventive nouns. This classification, guided by the parameters outlined in the Table 2.5, is instrumental in elucidating the intricate interplay between the noun's inherent semantic properties and the overall aspectual configuration of the LVC. The framework adopted herein posits that the stativity or eventivity of the noun significantly influences the temporal interpretation of the construction, shaping its durativity, dynamicity, telicity, and boundedness. Stative nouns, by their inherent nature, denote states of affairs that are relatively unchanging and lack dynamism. They typically describe conditions, qualities, or attributes that hold over a period of time. Consequently, when a stative noun is incorporated into an LVC, it tends to impart a sense of stability or continuation to the construction. For instance, an LVC containing a stative noun might express the maintenance of a particular state. In terms of the instrument, stative nouns would be characterized by a high degree of durativity [+durativity], but low dynamicity [-dynamicity]. Their contribution to telicity and boundedness is more variable and context-dependent, but they generally do not inherently impose telicity or boundedness on the event.

Table 2.6: Metric for evaluating the inherent temporal properties of a situation within noun of Indonesian LVCs.

Form of noun	Durativity	Dynamicity	Telicity	Boundedness
….	[+/-durativity]	[+/-dynamicity]	[+/-telicity]	[+/-boundedness]
….	[+/-durativity]	[+/-dynamicity]	[+/-telicity]	[+/-boundedness]

Note:
+ : presence
- : absence

Eventive nouns, in contrast, denote occurrences, processes, or activities that unfold over time and involve change. They describe dynamic situations with a clear sense of progression. When an eventive noun is used in an LVC, it introduces a sense of dynamism and potential change. The LVC, in this case, often describes the performance of an action or the occurrence of an event. According to the instrument of Table 2.6, eventive nouns are expected to exhibit a high degree of dynamicity [+dynamicity]. They may also contribute to telicity [+telicity] or boundedness [+boundedness], depending on the specific nature of the event. For example, a noun denoting a completed action would contribute both telicity and boundedness.

The analysis of noun element proceeds through a series of stage. First, each noun occurring within the LVCs identified in the corpus is categorized as either stative or eventive, based on established linguistic criteria and semantic intuitions, both are based on Table 2.5 as an instrument. This categorization is not always straightforward, as some nouns may exhibit characteristics of both categories, requiring careful consideration of the specific context of LVC. Second, the categorized nouns are then analyzed in conjunction with the verb element of the LVC, to determine how the noun's stativity or eventivity interacts with the verb's semantics. This involves examining whether the noun's inherent aspectual properties are reinforced or modified by the verb. The final stage is to create a categorical tabulation based on the analysis of the noun elements.

Moreover, machine-learning algorithms were implemented to analyze the relationships between verbs and nouns within LVCs. Specifically, the properties of these LVCs were quantified through the application of Random Forest Regression. The statistical software package XLSTAT (version 2020.4.1)[14] facilitated these analyses, employing aforementioned machine-learning methodologies. Throughout the analytic process, meticulous human-assisted validation was integrated to ensure the accuracy and consistency of the results. Furthermore, an iterative refinement strategy was adopted, enabling the continuous enhancement of our analytic methods and identification criteria in response to emergent patterns. To some extent, this method ensures a robust data foundation for the study and facilitated the effective attainment of the research objectives pertinent to the nuanced interplay of nouns and verbs in LVCs.

2.4 Author's relevant works

Table 2.7 provides an overview of all the relevant research papers published by the author. This information is specifically related to the research being conducted at present, making it an essential resource for anyone looking to gain a deeper understanding of the topic at hand. The table contains a brief of information, including the papers' titles, the publication dates, and other vital details relevant to the research in question. By consulting this table, one can better understand the author's previous work and how it relates to the current research project. Notably, the table also demonstrates a focus on Indonesian LVCs within the author's previous work. Several entries directly address specific verbs within LVCs, such as '*membawa*' (bring), '*memberi*' (give), '*memenuhi*' (meet), '*mengambil*' (take), and '*membuat*' (make). These entries, published in 2023 and 2024, indicate a sustained research interest in the morphosemantic and morphosyntactic properties of LVCs, which forms a foundation for the current thesis. Furthermore, the 2024 studies on Indonesian morphology, contribute to Chapters

[14] XLSTAT, a comprehensive statistical software package integrated within Microsoft Excel, proved invaluable in this study. Its user-friendly interface and diverse range of machine learning algorithms facilitated a multi-faceted analysis of verb behavior within LVCs. Specifically, XLSTAT's implementation of k-means clustering enabled the grouping of LVCs based on frequency patterns, while its Naïve Bayes classifier allowed for the probabilistic assessment of verb-noun collocations. Furthermore, the regression random forest functionality in XLSTAT provided insights into the relative importance of quantitative (e.g., frequency) and qualitative (e.g., grammatical distribution) features in predicting verb classification within LVCs. By leveraging these machine learning tools within XLSTAT, this research achieved a nuanced understanding of the complex interplay between lexical semantics and grammatical structure in the formation and interpretation of Indonesian LVCs.

1 and 2, suggesting their importance for the foundational and background sections of this thesis. In short, the organization of the table facilitates a clear understanding of the intellectual trajectory that informs the present research.

Table 2.7: Author's previous works.

Year	**Title**	**Relevance to chapter:**					
		I	**II**	**III**	**IV**	**V**	**VI**
2024	Towards an empirical understanding of *membawa* 'bring': Corpus insights into Indonesian light verb constructions			+	+	+	+
2024	A corpus-based study of *memberi* 'give' light verb constructions			+	+	+	+
2024	Morphosemantic features of *memenuhi* 'meet' in the light verb constructions of indonesian			+	+	+	+
2024	Analyzing prefix /me(n)-/ in the Indonesian affixation: a corpus-based morphology	+	+				
2024	A morphological analysis of the Indonesian suffixation: A look at the different types of affixes and their semantic changes	+	+				
2023	Morphosemantic features of *mengambil* 'take' in the light verb constructions of Indonesian			+	+	+	+
2023	Morphosyntactic features of *membuat* 'make' in the light verb constructions of Indonesian			+	+	+	+
2023	Morphosemantic features of *membuat* 'make' in the light verb constructions of Indonesian			+	+	+	+
2022	Identifying light verb constructions in Indonesian: A direct translation approach	+	+	+			

2.5 Ethical considerations

It is crucial to acknowledge the ethical considerations inherent in this research. These considerations encompass aspects of data privacy, informed consent, and responsible data usage. A primary concern lies in ensuring the privacy of individuals whose data is included within the corpus. This is particularly relevant since the part of corpus contains texts excerpts that might be identifiable. To mitigate this risk, anonymization techniques must be employed. This could involve rigorously removing personally identifiable information (PII) from the corpus data before analysis or utilizing anonymized pre-existing corpora constructed with ethical consideration in mind. Additionally, transparency about regarding the provenance and anonymization procedures employed for the corpus data used in this study is paramount.

Informed consent is another key ethical consideration. For large corpora, informed consent might not be feasible due to vast amount and potentially anonymous nature of the data. In this instance, reliance is placed on corpora that have been ethically constructed, ensuring informed consent was obtained at the point of data collection or that the data is anonymized to a degree that individual identification is not possible. Furthermore, responsible data usage is paramount. The corpus data employed for LVC research be used solely for the stated research objectives. Sharing or repurposing corpus data for unintended purposes raises ethical concerns. Adherence to any restrictions or access agreements associated with the corpus data and ensure all analysis comply with the ethical guidelines established by the corpus provider are essential.

It is also important to acknowledge potential biases within the corpus data itself. Corpora are often not perfect reflections of real-world language use. They might be skewed towards certain genre, registers, or historical periods. Awareness of these potential biases and consider how they might influence the findings related to LVC usage are crucial. Transparency about the corpus composition and potential limitations allows for a more nuanced interpretation of the research results. By carefully considering these ethical aspects–data privacy, informed consent, responsible data usage, and corpus bias–it can be ensured that this corpus-based exploration of LVCs in Indonesian is conducted in a responsible and ethical manner. This study, to some extent, contributes insights to the field of linguistics while respecting the privacy and rights of individuals whose language data is utilized in the research.

Chapter 3

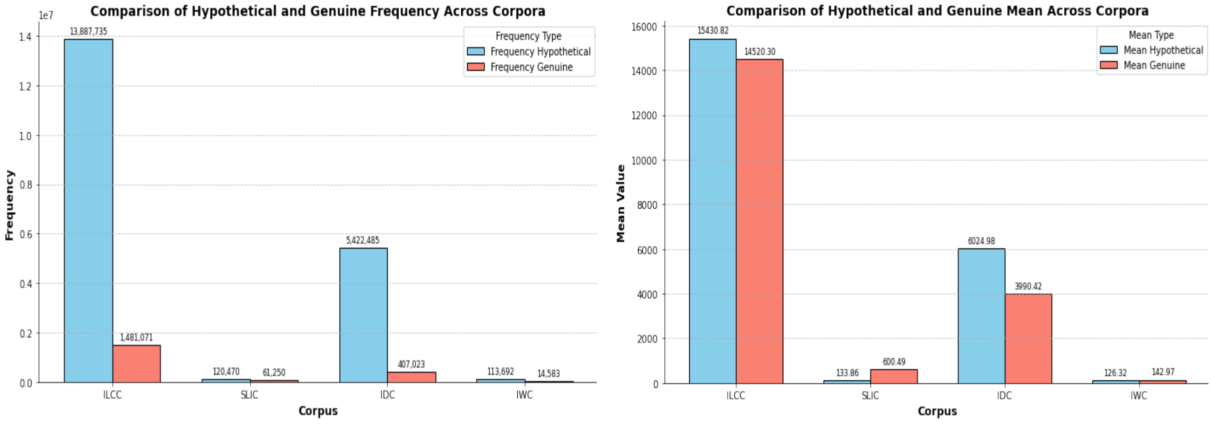
Frequency and distribution of Indonesian LVCs

3.0 Introduction

This chapter presents an analysis of Indonesian LVCs within four main corpora under investigation. The analysis seeks to elucidate the frequencies of these constructions within the investigated language corpora (§3.1). Furthermore, the analysis delves into the distributional patterns exhibited by these constructions (§3.2). Specifically, the analysis advances several central hypotheses as follows. Firstly, contrary to initial expectations of significant divergence due to inherent compositional disparities between the hypothetical and genuine LVC datasets, this study hypothesizes that specific linguistically well-defined LVC types will exhibit statistically significant correlations in frequency across the four corpora under investigation, when controlled for genre and time period (H1). This hypothesis seeks to ascertain whether observed frequency patterns transcend corpus-specific variations and reflect broader, systematic trends in LVC usage. Secondly, this study hypothesizes that the frequency distribution of LVCs, when analyzed as multiword expressions, will not strictly adhere to a Zipfian distribution, deviating significantly in the lower frequency ranges due to their complex syntactic and semantic properties (H2). This hypothesis challenges the assumption of uniform lexical distribution and explores the unique statistical behavior of LVCs as multiword units. Thirdly, this study will investigate whether naturally occurring distinctive clusters of LVCs exhibit statistically significant divergence in their distributional patterns across six predefined analytical dimensions. These dimensions are specifically designed to capture variations at the morphological, semantic, and syntactic levels (H3). Grounded in theoretical considerations concerning inherent linguistic features of LVCs and informed by empirical findings derived from clustering analyses of LVC frequency characteristics across all examined corpora, this hypothesis aims to elucidate the degree of (dis)connection between the identified clusters and their subsequent implications for a more nuanced understanding of Indonesian LVCs.

3.1 Frequency of Indonesian LVCs

To assemble the frequency of LVCs across the four corpora under investigation, a series of retrievals was executed, utilizing two distinct data collection instruments. The resultant identification outcomes, constituting a crucial preliminary finding, are presented in two discrete modalities: one derived from the hypothetical dataset (§3.1.1), and the other from the genuine dataset (§3.1.2). These matrices furnish a detailed account of LVC occurrence within the corpora. Subsequently, a K-means clustering analysis of the combined datasets, predicated on frequency homogeneity, was performed to delineate the natural distributional tendencies (§3.1.3). In addition, an in-depth analysis of Hypothesis 1 (H1) (§3.1.4) and Hypothesis 2 (H2) (§3.1.5) was also conducted to provide a more complete perspective on the distribution of LVCs within the four corpora analyzed.



(a) (b)

Figure 3.1: Presentation of the (a) total raw-frequency and (b) mean values of LVCs across the all investigated corpora based on hypothetical dataset and genuine dataset.

In general, Figure 3.1 provides a quantitative overview of the frequency of LVCs across four distinct corpora: ILCC, SLIC, IDC, and IWC. The figure reports two key types for each corpus: the raw-frequency of LVC occurrences and the mean-frequency per unit of corpus size.

The raw-frequency data reveals a striking disparity in the prevalence of LVCs across the corpora. Based on hypothetical dataset as illustrated in Figure 3.1(a), ILCC boats the highest raw-frequency, with a staggering 13,887,735 instances, orders of magnitude larger than the other three corpora. The IDC follows with 5,422,485 occurrences, while the SLIC and IWC exhibit considerably lower raw-frequencies of 120,470 and 113,692, respectively. This vast range suggests substantial difference in the size, composition, or LVC presence of these corpora. In addition, concerning frequency report of the genuine dataset, Figure 3.1(a) also offers a quantitative overview of the raw-frequency of LVCs across four distinct corpora. Examining the raw-frequency data, a clear disparity in the prevalence of LVCs is observed across the corpora. The ILCC again demonstrates the highest raw-frequency, with 1,481,071 instances, significantly exceeding the other corpora. The IDC follows with 407,023 occurrences. The SLIC and IWC exhibit considerably lower frequencies at 61,250 and 14,583, respectively. This wide range underscores potential differences in the size, composition, or LVC presence employed across these corpora.

Moreover, Figure 3.1(b) presents the mean frequency of LVCs per corpus unit. This matrix provides a normalized view of LVC prevalence, allowing for a more equitable comparison across corpora. Based on hypothetical dataset, the ILCC stands out with the highest mean frequency of 15,430.82 LVCs per unit, followed by the IDC at 6,024.98. The SLIC and IWC exhibit significantly lower mean frequencies of 133.86 and 126.32, respectively. These statistical observations highlight several key points. Firstly, the ILCC appears to be the most extensive corpus in terms of both raw size and density of LVCs. Secondly, the IDC, while considerably smaller than the ILCC, also contains a substantial number of LVCs. Finally, the SLIC and IWC, despite their relatively low raw frequencies, exhibit comparable mean frequencies, suggesting a similar degree of LVC representation within their respective sizes. B on the hypothetical dataset, the ILCC displays the highest mean frequency at 14,520.30 LVCs per unit. The IDC follows at 3,990.42. the SLIC and IWC demonstrate notably lower mean frequencies of 600.49 and 142.97, respectively. These statistical observations illuminate several key points. Firstly, the ILCC, even within the genuine dataset, remains the most extensive corpus in terms of both raw size and density of LVCs. Secondly, the IDC, though considerably smaller than the ILCC, still houses a substantial number of LVCs. Lastly, the SLIC and IWC, despite their lower raw-frequencies, exhibit a greater degree of comparability in their mean frequencies than in the hypothetical dataset, suggesting a more balanced representation of LVCs relative to their respective sizes within this genuine context.

3.1.1 Frequency details according to hypothetical dataset

Frequency serves as fundamental metric in the present corpus-based analysis. It functions as the primary quantitative indicator for gauging the prevalence – or conversely, the scarcity – of LVCs identified within the hypothetical dataset. By systematically tracking the frequency of LVC occurrences across the four corpora under investigation, the analysis gains empirical insights into their distributional patterns and relative prominence within the language. As illustrated in Figure 3.2, the pairwise matrix visualization[15] illustrates the interrelationships among four numerical variables – ILCC, SLIC, IDC, IWC – representing the *raw frequency* of LVC categories within the Hypothetical Dataset. Each data point in the constituent scatter plots depicts the paired values of two variables for a given LVC, thereby facilitating the observation of potential linear correlations, data point clustering, and the identification of outliers. The histograms situated along the diagonal provide a representation of the univariate distribution of each corpus. This analytical approach enables the expeditious identification of potential correlations, the dispersion of values, and the multifaceted nature of LVC distribution across corpora, offering both visual and intuitive insights.

Upon visual inspection of the pairwise scatter plot matrix, several initial observations regarding the relationships among variables can be discerned. The diagonal histograms reveal substantial disparities in the frequency distributions, with ILCC exhibiting a markedly higher magnitude compared to the other three variables. Examining the off-diagonal scatter plots, no strong linear correlations appear readily evident between any specific pair of variables. Instead, the point clouds suggest a diffuse pattern, indicating a lack of pronounced direct or inverse linear relationships in their raw frequencies within the Hypothetical Dataset. Furthermore, the

[15] The utilization of a pairwise matrix visualization offers several distinct advantages for illustrating the *raw frequency* of LVCs across a set of corpora. Firstly, this visual methodology facilitates the simultaneous examination of all possible bivariate relationships among the corpora, enabling a comprehensive assessment of inter-corpus similarities and dissimilarities in LVC frequency. Unlike univariate analyses that consider only one corpus at a time, the pairwise matrix allows for the direct comparison of frequency patterns between any two corpora, thereby revealing subtle nuances in LVC distribution that might otherwise be obscured. Secondly, the constituent scatter plots within the matrix provide a granular perspective on LVC behavior, representing each LVC as a discrete data point whose position reflects its frequency within the paired corpora. This representation allows for the identification of outliers, instances where specific LVC exhibit disproportionately high or low frequencies in one corpus relative to another, potentially indicative of corpus-specific biases or linguistic idiosyncrasies. Thirdly, the integration of diagonal histograms within the matrix furnishes a concise overview of the marginal frequency distributions for each corpus, complementing the bivariate comparisons with information about the internal variability of LVC frequencies within each individual corpus. This juxtaposition of bivariate relational data and univariate distributional summaries enhances the overall interpretability of the visualization, providing a more empirical understanding of LVC raw frequency patterns across the dataset.

presence of some scattered points at higher values in certain bivariate plots might suggest the occurrence of less frequent, yet potentially influential, co-occurrences across the different LVC categories.

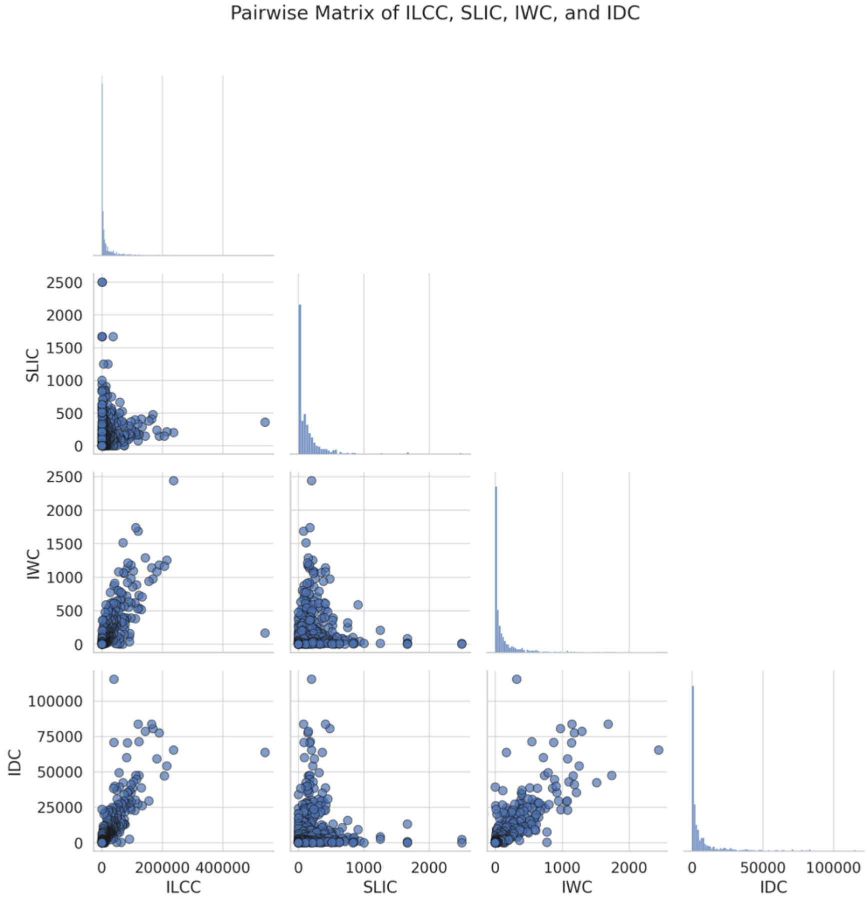


Figure 3.2: Pairwise matrix visualization of the relationships among four numeric variables—ILCC, SLIC, IDC, and IWC—that represent the *raw frequencies* of the LVCs categories in the Hypothetical Dataset.

Furthermore, Table 3.1 showcases sample list of the highest total-frequency (or Σ-frequency) LVCs extracted from a hypothetical dataset, offering a quantitative insight into their prominence within Indonesian language as reflected in the corpus. the LVCs listed exhibit a wide range of frequencies, with '*mencetak gol'* (to score a goal) leading with a Σ-frequency of 603,394. This is followed by a gradual decrease in Σ-frequency, ending in '*melakukan pemeriksaan*' (to inflict inspection) at 195,116 occurrences. Moreover, the standardized frequency (z-score) per unit of LVCs offers a normalized perspective, with '*mencetak gol*' (to score a goal) again leading at 13.189, and '*melakukan pemeriksaan*' (to inflict inspection) at the end of the list, with 39.39. The PMW across all LVCs within the Table 3.1 is diverse, from the highest level at 30321.578 to the latest list at 9800.039.

Table 3.1: Sample of highest Σ-frequency LVCs in hypothetical dataset.

No.	LVC	Total (Σ-Frequency)	Total (Z-Score)	PMW
1.	*mencetak gol* 'to score a goal'	603,694	13.189	30321.578
2.	*mengambil keputusan* 'take a decision'	305,520	6.439	15345.272
3.	*mengalami kesulitan* 'put difficulties'	270,296	5.461	13576.065
4.	*memberikan bantuan* 'to give help'	268,874	5.609	13504.663
5.	*melaksanakan tugas* 'do homework'	255,030	5.296	12809.324
6.	*membutuhkan waktu* 'require time'	250,737	5.198	12593.700
7.	*menghabiskan waktu* 'spends time'	249,208	5.164	12516.904
8.	*memberikan informasi* 'provides information'	224,450	4,603	11273.390
9.	*melakukan kegiatan* 'conducts activities'	205,734	4.179	10333.347
10.	*melakukan pemeriksaan* 'inflict inspection'	195,116	3.939	9800.039

In addition to highest Σ-frequency in hypothetical dataset, the current section provides sample of a medium Σ-frequency LVCs. Table 3.2 presents an overview of the sample of medium Σ-frequency LVCs extracted from a hypothetical dataset. The LVCs in this table exhibit a remarkably medium range of total frequencies, spanning from 36,878 for '*membuka jalan'* (pave the way) to 1,387 for '*memberi ciuman'* (give a kiss). Moreover, the standardized frequency (z-score) per unit of LVCs offers a normalized perspective, with '*membuka jalan'* (pave the way) again leading at 0.356, and '*memberi ciuman'* (give a kiss) at the end of the list, with -0.447. The PMW across all LVCs within the Table 3.2 is diverse, from the first level at 1852.261 to the tiniest list at 69.664.

Table 3.2: Sample of medium Σ-frequency LVCs in hypothetical dataset.

No.	LVC	Total (Σ-Frequency)	Total (Z-Score)	PMW
1.	*membuka jalan* 'pave the way'	36878	0.356	1852.261
2.	*mengulurkan tangan* 'give your hand'	23836	0.061	1197.204
3.	*memberikan keleluasaan* 'give escape'	17158	-0.090	861.790
4.	*melakukan orasi* 'give an oration'	13527	-0.172	679.417
5.	*melakukan audiensi* 'pay a state visit'	9631	-0.260	483.734

No.	LVC	Total (Σ-Frequency)	Total (Z-Score)	PMW
6.	*mencuri hati* 'steal someone's heart'	6493	-0.332	326.122
7.	*menjadi kesatuan* 'become united'	2286	-0.427	114.818
8.	*membuat tawaran* 'make an offer'	1396	-0.447	70.117
9.	*memberikan derma* 'give assistance'	1395	-0.447	70.066
10.	*memberi ciuman* 'give a kiss'	1387	-0.447	69.664

The last frequency band found in hypothetical dataset is low Σ-frequency. Table 3.3 provides a quantitative snapshot into the Σ-frequency of Indonesian LVCs occupying the low Σ-frequency stratum within a hypothetical dataset. The table showcases the ten fewest sample of low Σ-frequency LVCs identified, accompanied by their z-score relative to the entire corpus, and their PMW. The frequencies exhibited by these LVCs are notably sparse, ranging from a mere 1 to 9 occurrences. This aligns with their classification as ten fewest of the low Σ-frequency. The first list of LVC within this group, '*mengindahkan perhitungan*' (takes into account), appears only 9 times, while the remaining nine LVCs each occur between 1 and 4 times. This scarcity translates to z-score in relation to the entire corpus, all around to < 0 or -values.

Table 3.3: Sample of lowest Σ-frequency LVCs in hypothetical dataset.

No.	LVC	Total (Σ-Frequency)	Total (Z-Score)	PMW
1.	*mengindahkan perhitungan* 'takes into account'	9	-0.478	0.452
2.	*memberikan tonjokan* 'give a stroke'	8	-0.478	0.402
3.	*menyinari sesuatu* 'shine light on something'	7	-0.478	0.352
4.	*memberi pemicu* 'give boost'	4	-0.478	0.201
5.	*menjadi lipit* 'get pleated'	2	-0.478	0.100
6.	*memberikan dengusan* 'give a grunt'	2	-0.478	0.100
7.	*memulangkan sesuatu* 'return something'	1	-0.479	0.050
8.	*melakukan pelancongan* 'do a journey'	1	-0.479	0.050
9.	*memberikan tabokan* 'give a slap (in the face)'	1	-0.479	0.050
10.	*membuat koheren* 'make a coherent'	1	-0.479	0.050

This starkly contrasts with the high Σ-frequency LVCs discussed previously, highlighting the vast disparity in usage pattern across the Σ-frequency spectrum. In essence, Table 3.3 paints a picture of the sparsely populated realm of low Σ-frequency LVCs in the Indonesian hypothetical dataset. The extremely low frequency emphasizes the long tail nature of LVC distribution. The calculated metrics of z-score further illuminate the relative prominence of these LVCs within this low Σ-frequency band.

3.1.2 Frequency details according to genuine dataset

The exploration of Indonesian LVCs frequencies across the four selected corpora extends beyond the utilization of the hypothetical dataset. In this study, the genuine dataset also presents a valuable avenue for investigation, warranting careful consideration. Compiled through introspective elicitation from native Indonesian speakers, this genuine dataset comprises LVCs, whose existence and validity were further corroborated through raw frequency analysis within the aforementioned corpora. In this section, the findings of this initial examination will be systematically presented, offering a description of the raw frequency trends observed within LVCs in the genuine dataset.

As illustrated in Figure 3.3, the pairwise plot matrix for the Genuine Dataset reveals distinctive distributional characteristics and inter-variable relationship compared to the Genuine Dataset.[16] The diagonal histograms indicate that, similar to the hypothetical data, ILCC exhibits the highest frequency magnitude, albeit with a different distribution. SLIC and IWC show relatively low and concentrated frequency distributions, while IDC presents a wider spread, though still considerably lower than ILCC. Examining the bivariate scatter plots, the relationship between ILCC and the other variables (SLIC, IWC, and IDC) appears weak and scattered, suggesting minimal linear association in their genuine raw frequencies. Similarly, the plots of SLIC against IWC and IDC do not exhibit string linear patterns. However, a notable observation is the presence of clusters of points in several off-diagonal plots, particularly

[16] Notwithstanding the utility of the pairwise scatter plot matrix in providing an initial visual exploration of the raw frequency relationships, it inherently possesses limitations in capturing more nuanced distributional characteristics such as homogeneity or heterogeneity. The visualization primarily depicts raw co-occurrences, necessitating further analytical scrutiny to fully elucidate the underlying frequency patterns. Consequently, subsequent analyses should transcend the examination of mere raw frequencies and incorporate measures such as PMW to account for expected co-occurrence. Furthermore, the application of machine learning algorithms, e.g., K-means, holds promise in identifying inherent clusters within the frequency distributions across all examined corpora, irrespective of the dataset type, thereby enabling a more comprehensive understanding of the natural patterns.

between SLIC and IWC, and to some extent between IWC and IDC, concentrated at lower frequency values. This suggests that while strong linear correlations may be absent, there are instances where lower frequencies of these variables co-occur. The plot of IDC against ILCC and SLIC also shows a predominantly low-frequency clustering with some outliers extending to higher IDC values. Overall, the genuine frequencies appear to exhibit less pronounced linear relationships and more localized co-occurrence patterns at lower magnitudes compared to the hypothetical dataset.

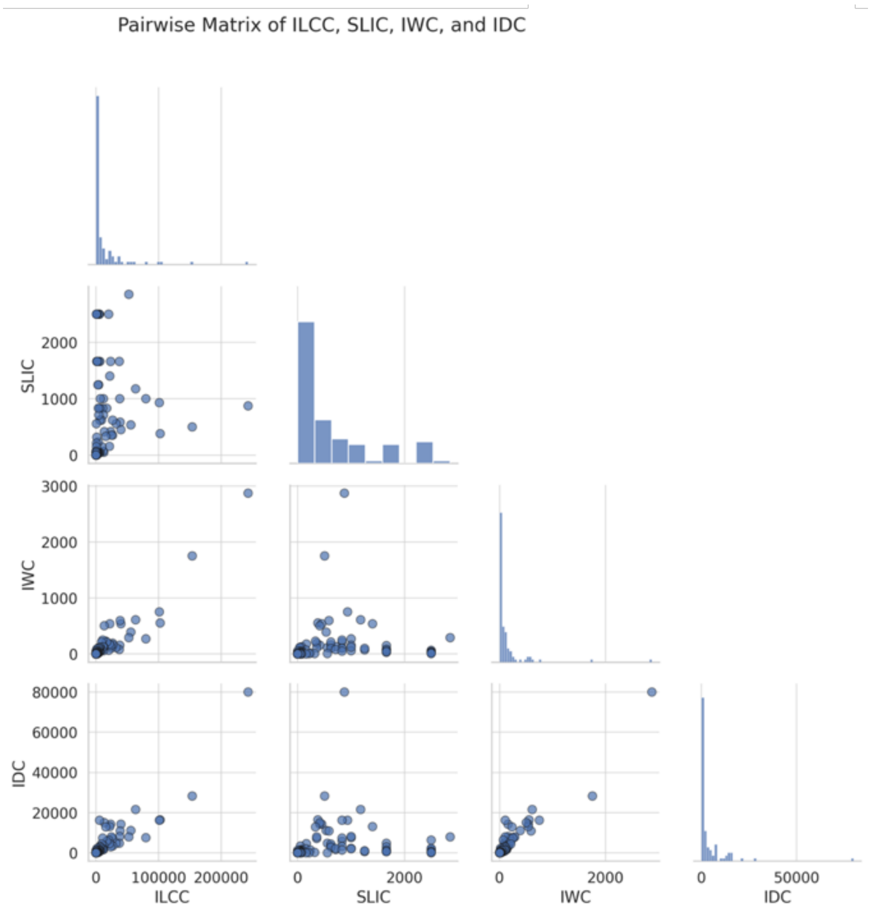


Figure 3.3: Pairwise matrix visualization of the relationships among four numeric variables—ILCC, SLIC, IDC, and IWC—that represent the raw frequencies of the LVCs categories in the Genuine Dataset.

Within the genuine dataset, raw frequency of LVCs can be seen in three distinct frequency bands based on their Σ-frequency: high Σ-frequency, medium Σ-frequency, and low Σ-frequency. Regarding the description for high Σ-frequency of LVCs within genuine dataset, Table 3.4 offers a quantitative glimpse into the frequency of Indonesian LVCs falling within the high Σ-frequency band of a genuine dataset. The table showcases the top 4 (four) LVCs within this band, accompanied by their total frequency across the corpus, their z-score relative to the entire corpus, and their PMW. The frequency exhibited by these high Σ-frequency LVCs span a considerable range, from 119,044 occurrences for '*bertolak belakang*' (to be contradictory) to 326,777 for '*masuk akal*' (to make sure). This range is also reflected in their z-score representation within the corpus, varying from 2.217 to 6.920. Additionally, the PMW column demonstrates a variance amongst these high Σ-frequency LVCs, from the first level at 16412.941 to the least list at 69.664.

Table 3.4: Sample of highest Σ-frequency LVCs in genuine dataset.

No.	LVC	Total (Σ-Frequency)	Total (Z-Score)	PMW
1.	*masuk akal* 'to make sense'	326,777	6.920	16412.941
2.	*campur tangan* 'to take action'	184,239	3.693	9253.723
3.	*pasang surut* 'ups and downs'	119,698	2,232	6012.040
4.	*bertolak belakang* 'to be contradictory'	119,044	2.217	5979.191

In addition to the first band of Σ-frequency, there is other tendency of Σ-frequency within the genuine LVCs dataset. Table 3.5 offers a quantitative glimpse into the circulation of LVCs within the medium Σ-frequency band of a genuine dataset. The frequencies exhibited by these medium Σ-frequency LVCs span a moderate range, from 37,662 occurrences for '*tertangkap basah*' (to be caught red-handed) to 88,477 for '*gulung tikar*' (to roll up the mat). This range is also reflected in their z-score representation within the corpus, varying from 0.374 to 1.525. The PMW across these LVCs is various, from 1891.639 to 4443.911.

Table 3.5: Sample of medium Σ-frequency LVCs in genuine dataset.

No.	LVC	Total (Σ-Frequency)	Total (Z-Score)	PMW
1.	*gulung tikar* 'to roll up the mat'	88,477	1.525	4443.911
2.	*turun tangan* 'to interfere'	86,421	1.478	4340.645
3.	*putus sekolah* 'to drop out of school'	67,957	1.060	3413.258
4.	*pandang bulu* 'to show favoritism'	63,777	0.965	3203.310
5.	*menutup mata* 'to close one's eyes'	54,686	0.760	2746.699
6.	*menjauhkan diri* 'to distance oneself'	50,579	0.667	2540.418

No.	LVC	Total (Σ-Frequency)	Total (Z-Score)	PMW
7.	*terbawa arus* 'to follow the crowd'	47442	0.596	2382.857
8.	*gigit jari* 'to bite one's fingers'	44010	0.518	2210.479
9.	*naik daun* 'to rise in popularity'	38380	0.390	1927.702
10.	*tertangkap basah* 'to be caught red-handed'	37662	0.374	1891.639

In addition to the two aforementioned bands of frequency, there is an evidence of exploration regarding the less frequent LVCs of genuine dataset. In Table 3.6, as anticipated for low Σ-frequency LVCs, the observed frequencies are notably sparse, ranging from a mere 1 to 118 occurrences.

Table 3.6: Sample of lowest Σ-frequency LVCs in genuine dataset.

No.	LVC	Total (Σ-Frequency)	Total (Z-Score)	PMW
1.	*bergoyang lidah* 'to gossip'	118	-0.476	5.927
2.	*dibanting harganya* 'to be sold at a very low price'	104	-0.476	5.224
3.	*tertusuk hatinya* 'to be deeply hurt'	92	-0.476	4.621
4.	*ada ekornya* 'there is more to it'	53	-0.477	2.662
5.	*pulang nama* 'to pass away'	38	-0.478	1.909
6.	*terkunci mulutnya* 'to be unable to speak'	19	-0.478	0.954
7.	*tertutup pikirannya* 'to be closed-minded'	12	-0.478	0.603
8.	*bertukar jalan* 'to switch places'	9	-0.478	0.452
9.	*berbalik kata* 'to go back on one's word'	1	-0.479	0.050
10.	*terbuka akalnya* 'to be open-minded'	1	-0.479	0.050

The example of LVC within this group, '*bergoyang lidah*' (to gossip) appears118 times and '*dibanting harganya*' (to be sold at a very low price) appears 104 times, while several other LVCs occur < 100 times—two samples are at lowest base, just one-time occurrence. This scarcity translates to negligible z-score in relation to the entire corpus, all rounding to minus value. The PMW column reveals a gradual decrease, ranging from 5.927 to 0.050. This emphasized the relatively small proportion of the corpus these low Σ-frequency LVCs represent.

3.1.3 Frequency groups from distinct natural clusters

The primary objective of this scrutiny is to employ K-means clustering[17] to discern and delineate homogeneous natural clusters within the dataset of 942 Indonesian LVCs, derived from the amalgamation of both hypothetical and genuine datasets. This clustering attempt is predicated upon the quantitative frequency data meticulously extracted from the four corpora under scrutiny. In detail, this experiment investigates the efficacy of K-means clustering in grouping Indonesian LVCs based on their frequency distribution across four corpora. The primary research question is whether K-means clustering can yield groupings that exhibit similar tendencies to those observed in frequency-based analysis. The expected output is a set of distinct clusters of Indonesian LVCs, categorized according to their inherent homogeneity characteristics. While K-means inherently involves binary decisions (assigning data points to clusters), the output is a multi-class classification, with the number of classes determined by the chosen value for K (the number of clusters).

The input feature for the K-means algorithm is the frequency of each LVC as extracted from the four aforementioned corpora. The current K-means clustering analysis in R-Studio was conducted using the following parameters: the optimal number of clusters (*k*) was determined using the three validation measures; the initialization method was employed for calculating distances between data points, as no compelling reasons were found to necessitate alternative distance measures; and the iteration settings, including the maximum number of iterations and convergence criteria, were adjusted to optimize the clustering process and ensure the stability of the resulting clusters. In detail, Figure 3.4 presents the cluster plot resulting from a K-means clustering analysis. In the context of analyzing LVCs, this visualization provides empirical insights into the central tendencies or representative characteristics of each identified cluster. It outlines three distinct clusters (marked as Q1, Q2, and Q3), each associated with a unique centroid as presented in Table 3.7. These centroids serve as reference points in a multi-dimensional space defined by the variables of LVCs frequency within ILCC, SLIC, IDC, and IWC. Their numerical values in each centroid offer insights into the distinguishing characteristics of each cluster.

[17] For more in-depth understanding of the theoretical underpinnings, the reader is directed to the following seminal works: Leski and Kotas (2018), McCarthy *et al.* (2016), Moisl (2010, 2010, 2015, 2021), Rosell (2009), and Thinsungnoen *et al.* (2015). Additional empirical evidence can be found in the studies conducted by: Al-Azzawy and Al-Rufaye (2017), Alkoffash (2012), Angheluta *et al.* (2004), Di and Gou (2018), Divjak and Fieller (2014), Everitt *et al.* (2011), Liu and Li (2010), and Naeem and Wumaier (2018).

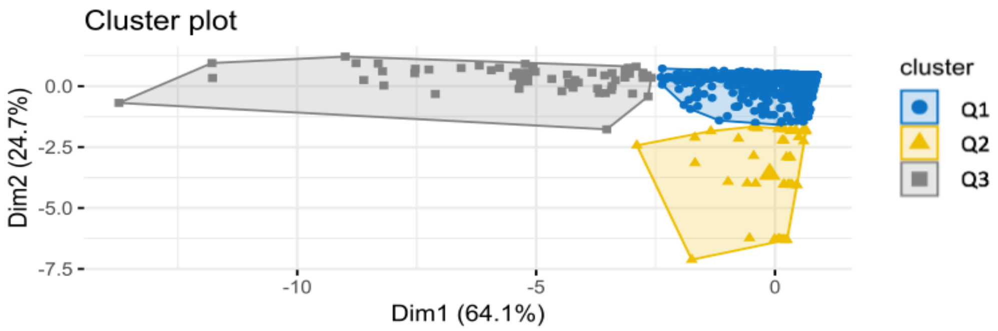


Figure 3.4: Cluster plot of the results of K-means clustering.

In Cluster Q1, the cluster is notable for its low negative values across all variable, i.e., ILCC (-0.19), SLIC (-0.21), IDC (-0.20), and IWC (-0.19). It suggests a grouping of LVCs sharing a pronounced tendency towards the linguistic properties associated with these variables. Based on these centroid coefficients, Q1 covers all low-frequency LVCs. In contrast, Cluster Q3 is characterized by a strikingly high value for all variable, i.e., ILCC (2.92), SLIC (0.15), IDC (3.20), and IWC (2.99). Grounded on these centroid coefficients, Q3 coverings entirely high-frequency LVCs. Lastly, cluster Q2, the cluster displays a relatively mixture profile. Three variable values are negative, i.e., ILCC (-0.08), IDC (-0.19), and IWC (-0.09); and one value is positive as in SLIC (3.60). Q2 is the middle cluster in the frequency distribution of LVCs in the four corpora analyzed.

Table 3.7: Centroid of clusters.

Cluster	ILCC	SLIC	IDC	IWC
Q1	-0.19	-0.21	-0.20	-0.19
Q2	-0.08	3.60	-0.19	-0.09
Q3	2.92	0.15	3.20	2.99

Specifically, the presented Figure 3.4 visualizes the distribution of data points across three clusters (color-coded blue, grey, and yellow) in a two-dimensional space defined by Dim1 and Dim2. These dimensions represent the first two principal components derived from a dimensionality reduction technique applied to the original, higher-dimensional feature space of the LVCs data. The percentages associated with each dimension (64.1% for Dim1 and 24.7% for Dim2) indicate the proportion of variance in the original data explained by these components. Visually, the three clusters exhibit a degree of separation, suggesting that the K-means algorithm has identified distinct groupings of LVCs based on their inherent PMW-frequency characteristics. Occupying the upper right quadrant, the blue cluster (Q1) appears relatively compact and well-defined. This compactness may indicate that the data within this cluster share strong similarities in their frequency features. The grey cluster (Q3), positioned in the upper left quadrant, exhibits a more dispersed distribution compared to the blue cluster. This dispersion suggests greater internal variability within this cluster regarding the frequency features under consideration. Located in the bottom-right portion of the plot, the yellow cluster (Q2) displays a specific spread between Q1 and Q3. This suggests that this cluster encompasses LVCs with a range of characteristics compared to the other two clusters. Some degree of overlap is observed between the clusters, particularly between the grey and yellow clusters. This overlap points to shared linguistic features among some verbs in these clusters, highlighting the complexity and potential gradience of linguistic categorization.

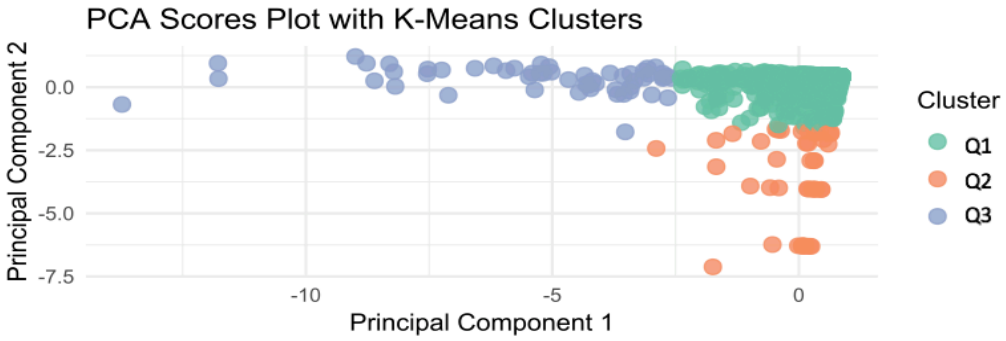


Figure 3.5: PCA Scores Plot of K-means clusters.

Additionally, according to subsequent analysis as illustrated in Figure 3.5, the presented plot visualizes the relationships between four linguistic corpora (ILCC, SLIC, IDC, and IWC) and the underlying dimensions (Dim1 and Dim2) derived from a Principal Component Analysis (PCA). In the context of analyzing inherent frequency features of LVCs within corpora, this plot aids in understanding how these corpora contribute to the variation in the non-linguistic features (PMW-frequency) observed across the identified clusters. The PCA has effectively reduced the dimensionality of the data, with Dim1 explaining 64.1% and Dim2 explaining 24.7% of the total variance. This indicates that these two dimensions capture a substantial portion of the variability in the non-linguistic features across the corpora. Table 3.8 reveals that the first two principles components (PC1:Dim1 and PC2:Dim2) account for a significant portion of the total variance (91.7%).

Table 3.8: Result of Principal Component Analysis.

Item	PC1	PC2	PC3	PC4
Standard deviation	1.6013	0.9945	0.51414	0.42716
Proportion of variance	0.6411	0.2472	0.06608	0.04562
Cumulative Proportion	0.6411	0.8883	0.95438	1.00000

Furthermore, the thresholding of LVCs into three empirically grounded clusters was based on standardized frequency-related variables, most notably *raw total frequency, per million words (PMW),* and z-scores derived from log-transformed frequency. This unsupervised method revealed a coherent stratification of the dataset into low-, medium-, and high-frequency clusters, each reflecting not only quantitative distinctions but also linguistic and functional divergence. Specifically, Cluster Q1 (low-frequency) encompasses LVCs with a PMW below 0.52, typically associated with z-scores less than -0.49, reflecting minimal recurrence in the dataset. Cluster Q2 (medium-frequency) includes constructions with a PMW ranging between 0.52 and 2.95, and corresponding z-scores roughly spanning -0.49 to +1.25. Lastly, Cluster 3 (high-frequency) comprises LVCs with PMW values exceeding 2.95 and z-scores greater than +1.25. Accordingly, the first cluster (Q1-low frequency) comprises 841 LVCs (*see* sample in Table 3.9), the second (Q2-medium frequency) encompasses 46 LVCs (*see* sample in Table 3.10), and the third (Q3-high frequency) contains 55 LVCs (*see* sample in Table 4.11).

Table 3.9: Sample list of LVCs within Cluster Q1 (low-frequency).

Twenty sample LVCs	
memberikan apresiasi ‘to commend outstanding ones’	*mengambil sikap* ‘takes a stand’
meluangkan waktu ‘take a moment’	*memberi contoh* ‘gives an example’
mengajukan pertanyaan ‘asks a question’	*memberikan nilai* ‘give a grade’
mengenyam pendidikan ‘to sit in the school’	*memberikan komentar* ‘give comments’
memberikan sambutan ‘give a speech’	*menjatuhkan sanksi* ‘impose sanctions’
mencari nafkah ‘make a living’	*menjalankan usaha* ‘run a small business’
menelan kekalahan ‘do losing’	*memberikan peluang* ‘provides an opportunity’
melancarkan serangan ‘launched an attack’	*memberikan izin* ‘give a permission’
memberikan jawaban ‘gives an answer’	*mengambil inisiatif* ‘take initiative’
melayangkan surat ‘to write a letter’	*memanjatkan doa* ‘make prayer’

In particular, Table 3.9 offers a window into the initial observation of distinctive characteristics in the first. This selection of twenty LVCs exemplifies the initial inspection of linguistic patterns and tendencies that characterize Cluster Q1 (low-frequency). This cluster represents the lowest frequency segment within the K-means-derived classification, and its rarity mirrored by distinctive morphological and semantic configurations. LVCs in this cluster predominantly exhibit affixed noun component, marking a morphological divergence from based-form constructions. These affixes are not merely decorative but encode syntactic constraints and semantic refinements, aligning with Mel'čuk's (2006) theory of linguistic signs where syntactic operate independently of form and meaning. The presence of affixation suggests that these LVCs are structurally marked, often realized through abstract nominalizations that are syntactically licensed but less frequently used in discourse. Besides, nouns in this cluster show a clear tendency towards [-material] profile, emphasizing abstraction. Examples include nouns like '*pendidikan*' (education) and '*sanksi*' (sanction), which lack physical referents and thus reinforce a conceptual orientation. This abstraction renders the constructions semantically elevated but less anchored in concrete discourse, which plausible contributes to their marginal frequency. From a formal standpoint, Cluster Q1 typifies the periphery of the LVC spectrum in frequency.

Table 3.10: Sample list of LVCs within Cluster Q2 (medium-frequency).

Twenty sample LVCs	
mengambil kesempatan ‘take a chance’	*membuka jalan* ‘pave the way’
memberikan keleluasaan ‘give escape’	*memutar otak* ‘to rack one’s brain’
melakukan orasi ‘give an oration’	*menyentuh hati* ‘to touch someone's heart’
melakukan penambahan ‘to make an addition’	*menyambung hidup* ‘to make a living’
melakukan audiensi ‘pay a state visit’	*membalas budi* ‘to return a favor’
membuat kegaduhan ‘create a scene’	*dimabuk cinta* ‘to be drunk in love’
melakukan penyergapan ‘make ambushes’	*mencuri hati* ‘to steal someone's heart’
membuat tawaran ‘make an offer’	*mengangkat sumpah* ‘to take an oath’
memberikan derma ‘give assistance’	*menjual diri* ‘to sell oneself’
memberi ciuman ‘give a kiss’	*membersihkan nama* ‘to clear one’s name’

Moreover, Table 3.10 presents a snapshot of the linguistic characteristics associated with the second cluster. Cluster Q2 occupies a middle frequency band within the LVC dataset, and its structural features position it between the extremes of affix-rich constructions and high-frequency idioms. LVCs in this cluster exhibit a more balanced distribution between *base* and *affixed* noun forms. This morphological equilibrium suggests a degree of flexibility in information, allowing these constructions to participate in both casual and formal registers. Additionally, this cluster reflects an almost split between [+material] (concrete) and [-material] (abstract) nouns. This morphological-feature diversity facilitates a wider range of semantic domains, from the physical (e.g., '*memberi ciuman*' (give a kiss)) to the conceptual (e.g., '*menyambung hidup*' (to make a living)). Such versatility implies that LVCs in Cluster Q2 can shift across discourse genres. Morphologically, this cluster represents a transitional zone, where affixation supports nuance without pushing the construction into structural rigidity. The relative accessibility of these forms may explain their mid-range frequency: they are not as formulaic as Cluster Q3, nor as marked as Cluster Q1. Rather, they function as semantically and morphologically adaptive units, ready to accommodate both informational and interactional needs.

Table 3.11: Sample list of LVCs within Cluster Q3 (high-frequency).

Twenty sample LVCs	
memberikan perhatian 'taking note of (something)'	*melakukan pemeriksaan* 'inflict inspection'
mengambil keputusan 'take a decision'	*mendapat dukungan* 'receives support'
mengalami kesulitan 'put difficulties'	*memberikan manfaat* 'to make use of something'
memberikan bantuan 'to give help'	*mengalami kerusakan* 'to be damaged'
memainkan peran 'plays a role'	*menjalani perawatan* 'receive treatment'
membutuhkan waktu 'require time'	*mengambil tindakan* 'take action'
menghabiskan waktu 'spends time'	*membuka peluang* 'make a chance'
masuk akal '*to make sense*'	*melakukan perjalanan* 'make a journey'
memberikan informasi 'provides information'	*mengambil langkah* 'takes a step'
melakukan kegiatan 'conducts activities'	*melakukan pengawasan* 'conduct supervision'

Lastly, Table 3.11 presents a sample selection of twenty LVCs that have been grouped together into the third cluster. Cluster 3 comprises the most frequent LVCs in the dataset, and its morphological profile reflects maximal accessibility and minimal complexity. Initially, these LVCs overwhelmingly contain *base-form nouns,* indicating that they avoid morphological embedding. This unmarked structure promotes rapid processing and usage across informal registers. The majority of noun elements are clearly [+material], signifying concrete referents such as people, objects, or actions. This high degree of material salience boosts conceptual immediacy, to some extent, allowing these LVCs to function effectively in both dialogic and didactic contexts. These LVCs often appear in routine expressions, idiomatic phrases, or fixed collocation (e.g., '*mengambil keputusan*' (take a decision), '*memainkan peran*' (plays a role), and '*masuk akal*' (to make sense)). The result is a set of LVCs that are both cognitively economical and communicatively efficient. Their formal brevity and semantic transparency place them at the core of usage-based grammar, where frequency and functional load reinforce each other. Morphologically, Cluster Q3 represents the most lexically unburdened zone of the system, which explains its prominence in both corpora and native speaker intuitions.

3.1.4 Cross-corpus frequency analysis of linguistically defined LVCs: Testing for Genre and Temporal Stability

After describing the frequency of LVCs based on the two datasets used in this study and collectively grouping them based on K-means clustering analysis, a deeper examination needs to be conducted to understand the inherently characteristic of frequency. To rigorously assess Hypothesis 1, which posits statistically significant correlations in the frequency of specific LVC types across the four corpora, this subsection presents a quantitative analysis employing statistical correlation tests. The analysis explores the consistency of LVC type distribution across the ILCC, SLIC, IDC, and IWC corpora. Central to this investigation is the categorization of LVCs into four linguistically well-defined types: (a) true light verb constructions, (b) vague action verb constructions, (c) eventive light verb constructions, and (d) stative light verb constructions. By calculating correlation coefficients for each LVC type, the study aims to determine the extent to which their occurrence patterns are consistent across corpora, thereby revealing potential underlying trends in LVC usage that transcend corpus-specific idiosyncrasies. Given that all corpora are text-based and collected from online language-usage contexts, the influence of genre is controlled, allowing for a focused examination of temporal factors. However, the time periods of the corpora vary, with ILCC collected in 2013, SLIC in 2010, IDC in 2020, and IWC in 2012. Therefore, the analysis will also consider the potential influence of time period on these correlations, where feasible. Specifically, we will explore whether the observed correlations are affected by the temporal distance between the corpora, potentially revealing trends in LVC usage over time.

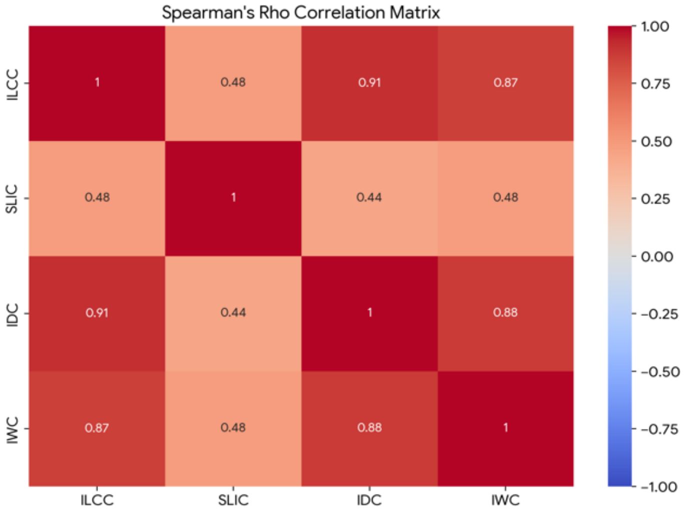


Figure 3.6: Spearman's rank correlation to discern pattern in the distribution of LVCs across the four corpora.

The interrelationships among the lexical frequency measures were examined via Spearman's rank correlation to discern pattern in the distribution of LVCs across the four corpora.[18] As illustrated in Figure 3.6, the analysis revealed robust positive correlations between several variables, suggesting a degree of systematicity in LVC usage. Specifically, a strong positive association was observed between ILCC and IDC (r_s = .91, p < .001, N = 942), indicating that the relative frequencies of LVCs and other lexical items in these two corpora exhibit a high degree of concordance. This suggests that LVCs, along with other lexical components, tend to occur with similar relative frequencies in both ILCC and IDC. A similarly strong positive correlation was found between ILCC and IWC (r_s = .87, p < .001, N = 942), further supporting the notion that LVC usage patterns are broadly consistent between these corpora. Furthermore, IDC and IWC demonstrated a very strong positive correlation (r_s = .88, p < .001, N = 942), suggesting that these two corpora share a particularly high degree of similarity in their lexical distributions, implying a shared preference or constraint in LVC selection and frequency. These findings, taken together, suggest a substantial degree of covariation in the occurrence of lexical items, including LVCs, across these three corpora,

[18] *See* Kilgarriff (1997, 2001) and Kilgarriff and Rose (1998).

implying that common underlying factors, such as shared linguistic development or similar stylistic preferences, may influence their lexical composition.

In contrast, SLIC exhibited only moderate positive correlations with ILCC ($r_s = .48$, $p < .001$, $N = 942$), IDC ($r_s = .44$, $p < .001$, $N = 942$), and IWC ($r_s = .48$, $p < .001$, $N = 942$), indicating a comparatively weaker relationship. This disparity warrants further scrutiny, particularly in the light of the temporal differences between the corpora. While the controlled genre (text-based online language usage) mitigates potential confounding influences of register variation, the collection years vary (ILCC: 2013, SLIC: 2010, IDC: 2020, IWC: 2012). Consequently, the lower correlations involving SLIC may, at least in part, reflect diachronic changes in language usage, potentially indicating a shift in the frequency or distribution of specific LVCs over time. The earlier collection date of SLIC (2010) may account for its relatively distinct lexical profile compared to the more recent corpora, suggesting that certain LVCs may have increased or decreased in prevalence in the intervening years. However, it is crucial to acknowledge that the presence or absence of specific LVCs is not directly measured by correlation coefficients, which instead reflect the degree of association in overall lexical frequency. Therefore, the observed correlations point to broader shifts in lexical patterns, within which LVCs are embedded.

Furthermore, the subsequent analysis aims to understand the behavior of LVCs across different corpora. Specifically, it involved in whether the way LVCs are used is consistent or variable across the ILCC, SLIC, IDC, and IWC corpora. This has implications for understanding how LVCs evolve, whether they are influenced by context, and what underlying linguistic principles govern their use. In detail, Hypothesis 1 focuses on the consistency of LVC distribution. It posits that there should be statistically significant correlations in the frequency of specific LVC types across the corpora, i.e., the categorization of LVCs into four linguistically well-defined types: (a) true light verb constructions, (b) vague action verb constructions, (c) eventive light verb constructions, and (d) stative light verb constructions; where (a) is coded `'tl'` and (b) is coded `'va'` which are components of `TYPE_VERB_1`, and (c) is coded `'St'` and (d) is coded `'Ev'` which are components of `TYPE_VERB_2`. This implies that if a particular type of LVC is common in one corpus, it should be relatively common in the other corpora.

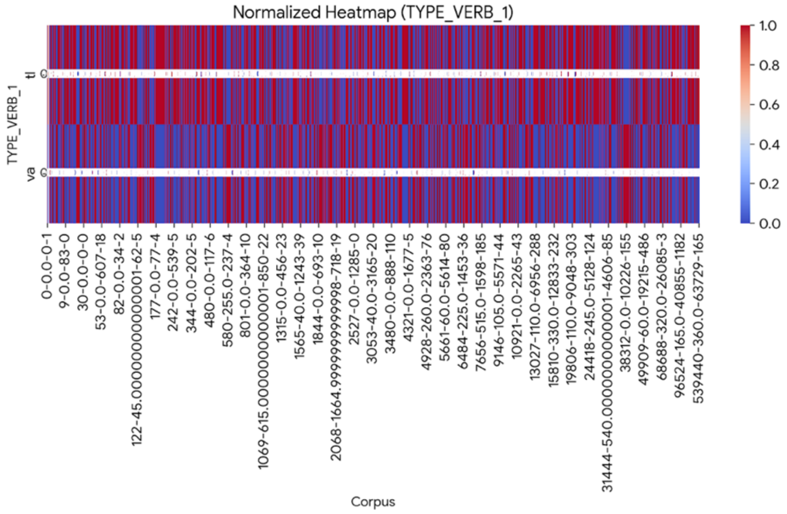


Figure 3.7: Normalized heatmap for `TYPE_VERB_1` distribution across the four corpora.

As the analysis strategy, the initial plan was to use *correlation coefficients* to measure the consistency. Correlation would show if the frequency of a give LVC type in one corpus is positively related to its frequency in other corpora. However, the data structure (categorial LVC types) led to the use of *Chi-Square tests of independence.* Chi-Square tests assess whether there is an *association* between the LVC type and the corpus. In this context, a *lack* of association (failing to reject the null hypothesis) supports the idea of *consistency* because it means the proportions of LVC type are similar across corpora. As for the analysis of `TYPE_VERB_1`, the Chi-Square test revealed no significant association between `TYPE_VERB_1` and the corpora (χ^2 = 936.68, df = 936, p = .488). The heatmap, as in Figure 3.7, also shows the proportions of '`tl`' and '`va`' within each corpus. The relatively uniform colors visually support the Chi-Square test results, indicating similar proportions across corpora. Therefore, the Chi-Square test and the heatmap consistently demonstrate that the distribution of LVCs categorized by is `TYPE_VERB_1` consistent across the four corpora. The proportion of '`tl`' and '`va`' do not vary significantly from ILCC to SLIC to IDC to IWC. Accordingly, some linguistic implications emerged. If '`tl`' and '`va`' represent distinct linguistic classes of LVCs, this consistency might suggest that the factors governing the choice between these classes are stable across different online language contexts. It could imply that corpus-specific factors, i.e., time consideration, do not strongly influence the relative usage of these LVC types. To be more precise, the results align with the general idea of Hypothesis 1 (consistency of LVC type distribution), even though we used a Chi-Square test instead of correlation. The *lack* of a significant association in the Chi-Square test is the statistical evidence supporting the *consistency* of the distribution.

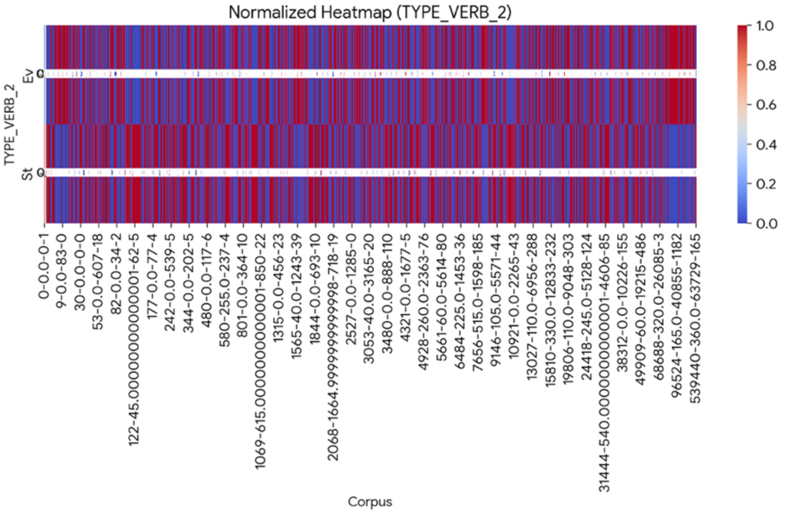


Figure 3.8: Normalized heatmap for `TYPE_VERB_2` distribution across the four corpora.

As for the analysis of `TYPE_VERB_2`, the Chi-Square test revealed no significant association between `TYPE_VERB_2` and the corpora (χ^2 = 935.194, df = 936, p = .501). The heatmap shows the proportions of 'St' and 'Ev' within each corpus. The relatively uniform colors visually support the Chi-Square test results, as in Figure 3.8, indicating similar proportions across corpora. The Chi-Square test and the heatmap consistently demonstrate that the distribution of LVCs categorized by `TYPE_VERB_2` is consistent across the four corpora. The proportions of 'St' and 'Ev' LVCs do not vary significantly from ILCC to SLIC to IDC to IWC. The observed consistency in the distribution of eventive 'St' and 'Ev' LVCs across the corpora carries significant linguistic implications. Specifically, if these categories represent genuinely distinct linguistic classes, their stable proportions suggest that the factors governing their selection are robust across diverse online language contexts. This, in turn, implies that corpus-specific variables, such as time-period, exert minimal influence on the relative frequencies of these LVC types. Crucially, these findings directly align with the overarching premise of Hypothesis 1, which posits consistent occurrence patterns of LVC types. Although the analytical methodology employed a Chi-Square test of independence rather than the initially proposed correlational analysis, the *absence* of statistically significant association serves as compelling evidence for the hypothesized consistency in LVC type distribution across the corpora.

In summary, the Chi-Square tests, conducted to evaluate the consistency of LVC type distribution across the corpora, revealed no statistically significant association between either LVC type categorization and the corpora. This *failure to reject* the null hypothesis suggests that the proportions of these LVC types are relatively consistent across the ILCC, SLIC, IDC, and

IWC corpora. In the context of Hypothesis 1, which posits statistically significant correlations (implying consistent occurrence patterns), the Chi-Square test results indicate that the distribution of LVC types is *not* significantly influenced by the corpus. This finding lends support to the notion that the occurrence patterns of these LVC types are indeed consistent across the corpora, as originally hypothesized, though this consistency is demonstrated through the *absence* of a statistically significant association in the Chi-Square test, rather than a strong correlation. This consistency in LVC type distribution across corpora, despite differences in collecting dates and specific content, may suggest that the categorization of LVCs into these types reflects fundamental linguistic properties that are relatively stable in online language use. It implies that these LVC types are not strongly influenced by temporal changes or corpus specific biases, supporting the idea of underlying trends in LVC usage that transcend individual corpora.

3.1.5 Deviations from Zipfian Distribution in the frequency profiles of LVCs as multiword expressions

The second hypothesis (H2) predicts a *deviation* from a strict Zipfian distribution[19], especially in the *lower frequency ranges.* This deviation is attributed to the *complex syntactic* and semantic *properties* of LVCs as multiword expressions. The hypothesis challenges the idea of a uniform distribution of lexical items and highlights the unique statistical behavior of multiword units. Specifically, Zipf's Law states that in a large corpus of text, the frequency of a word is inversely proportional to its rank in the frequency table. The most frequent word occurs about twice as often as the second most frequent word, three time as often as the third most frequent word, and so on. If *f* is the frequency of a word and *r* is its rank, then Zipf's Law can be roughly expressed as follows:

$$f \propto \frac{1}{r} \; or \, f \approx \frac{K}{r} \quad \text{where K is constant.} \qquad (3.1)$$

As the linguistic significance, Zipf's Law is observed in many languages and text types. However, there is no single perfect statistical test to definitely "prove" Zipf's Law, as it is more of an observed tendency. Accordingly, we can use several approaches to assess how well the

[19] *See* Altmann (1985, 1997, 2025), Köhler *et al.*, (2005), Oakes (2019), and (Zipf, 2013).

data fits a Zipfian distribution, such as *Visual Inspection (Log-Log Plot)*, *Regression Analysis,* and *Goodness-of-Fit Tests.* These types of analysis are presented sequentially in the following sections. First, Log-Log Plot Analysis. The data was loaded form “Ch.3-3.1.4–Sheet3.csv” using `Pandas`. The `DataFrame dF` has columns like LVC, ILCC, SLIC, IDW, IWC, TOTAL, and RANK. A log-log scatter plot was generated using `Matplotlib` and `Altair`. Both axes (x: `RANK`, y: `TOTAL`) were transformed to a logarithmic scale. A theoretical Zipf’s Law line was added to the plot. This line represents the expected relationship if the data followed Zipf’s Law perfectly (a straight line with a slope of -1 on the log-log scale). The frequency of the most frequent LVC was used to scale the theoretical Zipf’s Law line.

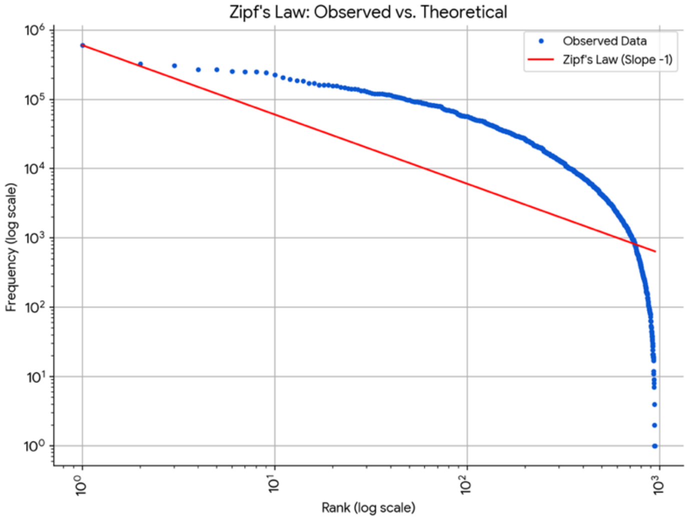


Figure 3.9: Result of log-log plot analysis on observed vs. theoretical LVCs (Zipfian distribution).

The plot reveals a general downward trend, as demonstrated in Figure 3.9, indicating an inverse relationship between the rank and the frequency of LVCs. This aligns with the basic principle of Zipf’s Law: more frequent LVCs tend to have higher ranks, and vice-versa. The observed data points show a *roughly linear* pattern, suggesting that Zipf’s Law captures a significant aspect of the frequency distribution of LVCs. While the overall trend is linear-like, there are noticeable deviations from a perfect straight line, particularly at the lower frequency (high rank) end of the plot. The deviations of the observed data points from the line highlight the ways in which the actual frequency distribution of LVCs differs from the theoretical Zipfian distribution. Moreover, for the high-frequency end (low rank), the data points representing the most frequent LVCs (those with low ranks, towards the left of the plot) tend to align more closely with the theoretical Zipf’s Law line (red line). This suggests that Zipf’s Law provides a relatively good fit for the most common LVCs. Additionally, for the low-frequency end (high rank), the data points representing the less frequent LVCs (those with high ranks, towards the right of the plot) show a more pronounced deviation from the theoretical line. Specifically, there is a clear mark for increased scatter and slight upward curvature. As the increased scatter, there is more variability in the frequency of low-frequency LVCs compared to high-frequency LVCs. The points are more dispersed and do not adhere as closely to a linear trend. As the slight upward curvature, the observed data points tend to be located *above* the theoretical Zipf's Law line in this region. This indicates that the low-frequency LVCs are somewhat *more frequent* than predicted by a strict Zipfian distribution. In other words, there are more low-frequency LVCs than expected if Zipf's Law held perfectly.

Therefore, Hypothesis 2 (as previously stated) predicted that the frequency distribution of LVCs would *not* strictly adhere to a Zipfian distribution, especially in the lower frequency ranges. The findings from the log-log plot *support* Hypothesis 2. The observed deviations from linearity, particularly the increased scatter and upward curvature in the low-frequency region, demonstrate that the distribution of LVCs is not perfectly Zipfian. The deviations from Zipf's Law suggest that factors beyond simple rank-based frequency influence the occurrence of LVCs, especially the less frequent ones. These factors are related to the linguistic nature of LVCs as multiword expressions. For example: idiomaticity—some LVCs have non-compositional meanings, making their occurrences less predictable; syntactic constraints—LVCs may have specific syntactic requirements that limit their distribution; and semantic specificity—LVCs often express specialized meanings, leading to restricted usage contexts.

Second, Regression Analysis. The analysis aims to statistically model the relationship between the logarithm of rank and the logarithm of frequency, to obtain a precise estimate of the slope of this relationship (a slope of -1 is expected if Zipf's Law holds perfectly), and to assess the overall fit of the model using the R-squared value. As a method, this analysis uses ordinary least squares (OLS) linear regression on the log-transformed data. This is a standard technique for analyzing power-law relationship like Zip's Law. As illustrated in Figure 3.10, the regression analysis, employing OLS to predict the logarithm of total frequency (`log_total`), revealed a statistically significant relationship with the logarithm of rank (`log_rank`) ($p < .001$). The model, explaining 68.5% of the variance in `log_total` (R-squared = 0.685), indicated that about 68.5% of the variation in the log-transformed frequency of LVCs can be attributed to its linear relationship with log-transformed rank. However, the crucial finding is the slope coefficient for `log_rank`, which was -1.9075. This value deviates substantially from the -1.0-slope predicted by Zipf's Law, suggesting that the logarithm of total frequency decreases 1.9075 units for every one-unit increase in the logarithm of rank, a steeper decline than anticipated. This result is precise, as evidenced by the small standard error of the slope coefficient (0.042), and the 95% confidence interval for the slope, ranging from -1.990 to -1.825 and excluding -1, further reinforces the significant departure from a strict Zipfian distribution. The y-intercept of the regression line was 19.4283, a value less directly relevant to Zipf's Law interpretation.

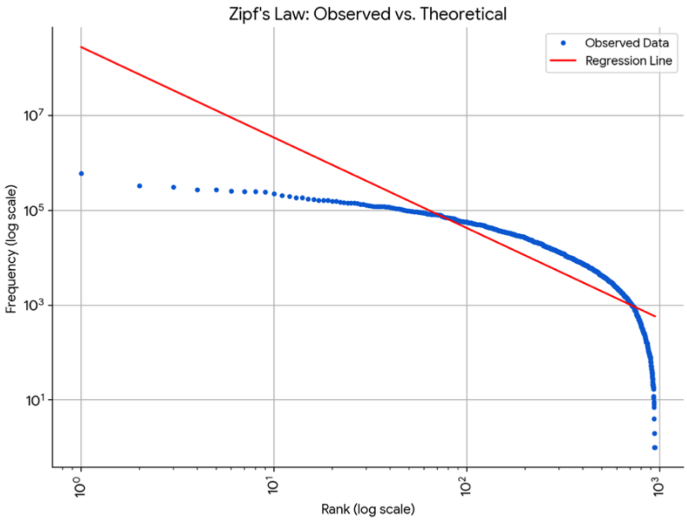


Figure 3.10: Result of regression analysis on observed vs. theoretical LVCs (Zipfian distribution).

Summary of regression analysis encompass several findings, i.e., significant relationship, deviation from Zipf's Law, and moderate fit. There is a highly statistically significant linear relationship between the logarithm of rank and the logarithm of total frequency. The slope of regression line (-1.9075) is considerably steeper than the -1.0 predicted by Zipf's Law, indicating a clear deviation from a strict Zipfian distribution. The R-squared value (0.685) suggest that the regression model explain a moderate amount of the variance in the log-transformed frequency data. In the context of Hypothesis 2, it predicted a deviation from a strict Zipfian distribution, particularly in the lower frequency ranges, due to the complex syntactic and semantic properties of LVCs. The regression analysis strongly supports this hypothesis. The slope of -1.9075 confirms that the relationship between rank and frequency is significantly different from what Zipf's Law predicts. The more rapid decrease in frequency with increasing rank suggests that lower-frequency LVCs are even less frequent than expected under a pure

Zipfian model. The moderate R-squared value indicates that while rank is a strong predictor of frequency, other linguistic factors are also influencing the distribution of LVCs.

Moreover, the residual analysis can proceed with analyzing the residuals from the regression to understand the deviations from Zipf's Law in more detail. Residuals are the differences between the observed values (actual log frequencies) and the values predicted by the regression model. The analysis has several purposes, such as: to identify systematic patterns in the deviations from Zipf's Law that might not be apparent from the overall regression results and to see if there is a trend in the residuals as rank increases (i.e., as the analysis more towards lower-frequency LVCs). As illustrated in Figure 3.11, X-axis is Rank (log scale)—this axis represents the rank of the LVCs, plotted on a logarithmic scale. This allow us to visualize the residuals across the entire range of ranks, from the most frequent (left side) to the least frequent (right side). Y-axis is Residuals (Observed – Predicted Log (Frequency))—this axis represents the residuals, which are the differences between the observed logarithm of the total frequency and the logarithm of the frequency predicted by the regression model. Some indicators are as follows: a residual of 0 means the model perfectly predicted the frequency; a positive residual means the model *underestimated* the frequency (the actual frequency was higher than predicted); and a negative residual means the model *overestimated* the frequency (the actual frequency was lower than predicted). If the regression model (and Zipf's Law) perfectly captured relationship between rank and frequency, the residuals should be randomly scattered around the horizontal line at y = 0. There should be no clear patterns or trends.

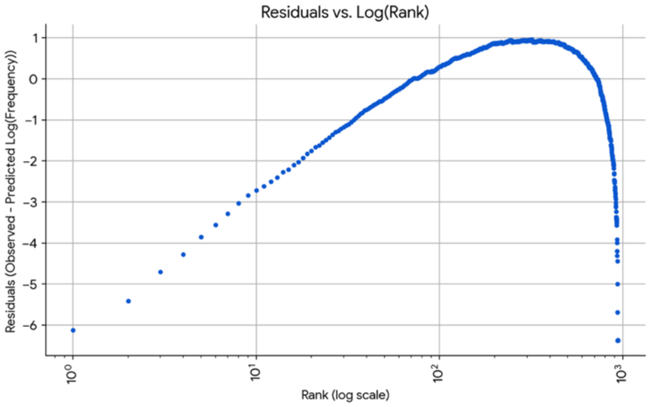


Figure 3.11: Result of residual analysis.

As the overall trend, the residuals do *not* appear to be randomly scattered. Instead, there is noticeable trend. At the left side of the plot (low ranks, high-frequency LVCs), the residuals are relatively small and scattered around zero. This indicates that the regression model fits the high-frequency LVCs reasonably well. As we move to the right side of the plot (high ranks, low-frequency LVCs), the residuals tend to become more *positive.* This means that the regression model *underestimates* the frequency of the low-frequency LVCs. In other words, the low-frequency LVCs occur more often than predicted by the regression model (and by a strict Zipfian relationship). As for curvature, there might be a subtle curvature in the distribution of the residuals seems to be slightly larger at the high-rank end compared to the low-rank end. This suggests that the model's predictions are less precise for low-frequency LVCs.

Therefore, several linguistic implications are as follows. The systematic pattern of *positive residuals* at the high-rank end has important linguistic implications. It indicates that the Zipfian model, which is based purely on rank, *underestimates* the occurrence of less frequent LVCs. This underestimation is likely due to the linguistic properties of LVCs, which are not captured by simple rank-based frequency. Low-frequency LVCs may be more idiomatic, meaning their meaning is not predictable from their components. This could make them more likely to appear in specific contexts, boosting their frequency. Concerning syntactic constraints, some low-frequency LVCs might have specific requirements that limit their overall occurrence but increase their likelihood in certain constructions. Regarding semantic specificity, less frequent LVCs might express very specific meanings, making them essential in particular discourse contexts. In short, the residual analysis provides further evidence against a strict Zipfian distribution for LVCs and supports Hypothesis 2, which predicted deviations, especially in the lower frequency ranges.

Third, Goodness-of-Fit Tests. Goodness-of-fit tests assess how well an observed distribution matches a theoretical distribution. In this case, the analysis wants to see how well the observed frequency distribution of LVCs matches a theoretical Zipfian distribution. Traditional goodness-of-fits test can be done with Kolmogorov-Smirnov (K-S) test. The K-S test compares the cumulative distribution function (CDF) of the observed data to the CDF of a specified theoretical distribution. The K-S test statistic measures the maximum distance between the two CDFs. A small p-value (< 0.05) suggests a significant difference between the observed and theoretical distributions. The K-S test yielded a KS statistic of 0.596854 ($p = 5.11e-321$), indicative of statistically significant divergence between the CDF of the observed normalized ranks and the theoretical Zipfian CDF. This KS statistic, representing the maximum dissimilarity between the two CDFs, provides a quantitative measure of the departure from a strict Zipfian distribution. However, in cognizance of the inherent limitations of applying such test to complex linguistic phenomena, and acknowledging the expected deviations, particularly with multiword expressions like LVCs, this result primarily serves to corroborate the findings of the preceding log-log plot and regression analyses.

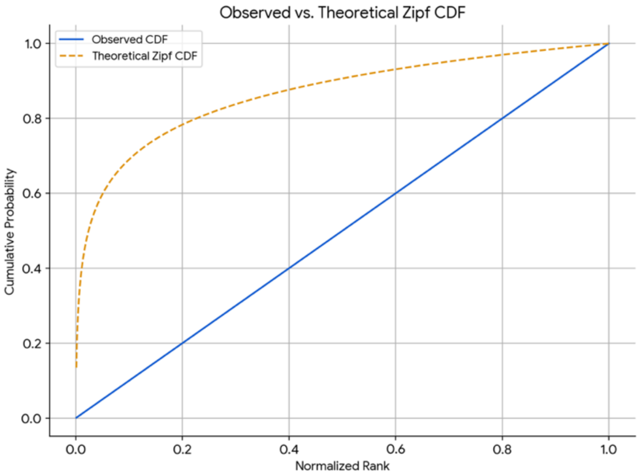


Figure 3.12: Result of goodness-of-fit tests.

Moreover, to enhanced understanding the K-S test, a visualization of CDF Plot is required. As illustrated in Figure 3.12, X-axis is normalized rank. This axis represents the rank of the LVCs, normalized to a range between 0 and 1. 0 corresponds to the highest-ranked (most frequent) LVC, and 1 corresponds to lowest rank (least frequent) LVC. Y-axis is a cumulative probability. This axis represents the cumulative probability. For a given normalized rank value on the x-axis, the y-axis shows the proportion of LVCs with a normalized rank less than or equal to that value. Solid-blue line is observed CDF. This line shows the cumulative distribution of the normalized ranks in the actual data. It represents the empirical distribution of LVC ranks. Dashed-orange line is theoretical Zipf CDF. This line shows the cumulative distribution that would be expected under perfect Zipfian distribution. It is calculated based on the `zipf_cdf` function defined for this calculation.

As the overall comparison, the plot shows that the observed CDF and theoretical Zipf CDF are *not identical*, indicating that the observed distribution of LVC ranks deviates from a perfect Zipfian distribution. This confirms the results of the log-log plot and regression analysis. Furthermore, there is specific deviations as shown in Figure 3.12. at the beginning of the plot (low normalized ranks, high-frequency LVCs), the observed CDF and the theoretical CDF are relatively close. This suggests that Zipf's Law provides a reasonable approximation for the most frequent LVCs. The most pronounced difference between the two CDFs occurs toward the right side of the plot (high normalized ranks, low-frequency LVCs). In this region, the observed CDF is consistently *above* the theoretical Zipf CDF. Accordingly, the fact that the observed CDF is above the theoretical CDF at higher ranks means that, in the data, there is a *higher proportion* of LVCs with relatively *low* normalized ranks (i.e., less frequent LVCs) than what a perfect Zipfian distribution would predict. This observation aligns with earlier findings from the log-log plot, where we saw that the low-frequency LVCs were more frequent than expected under a strict Zipfian model. In other words, the CDF plot visually reinforces the conclusions drawn from the log-log plot and regression analysis. All three analyses point to a deviation from a strict Zipfian distribution, particularly for the less frequent LVCs. The K-S statistic (0.596854) and the extremely small p-value (5.11e-321) from the K-S test quantify the overall difference between the observed and theoretical CDFs. It clearly illustrates how the cumulative distribution of LVC ranks in the data differs from the theoretical Zipfian distribution, especially for the less frequent LVCs. This visualization strengthens the evidence for rejecting a strict Zipfian model for LVC frequency distribution and support the hypothesis that linguistic factors play a significant role. As mentioned before, these factors include idiomaticity, syntactic constraints, and semantic specificity.

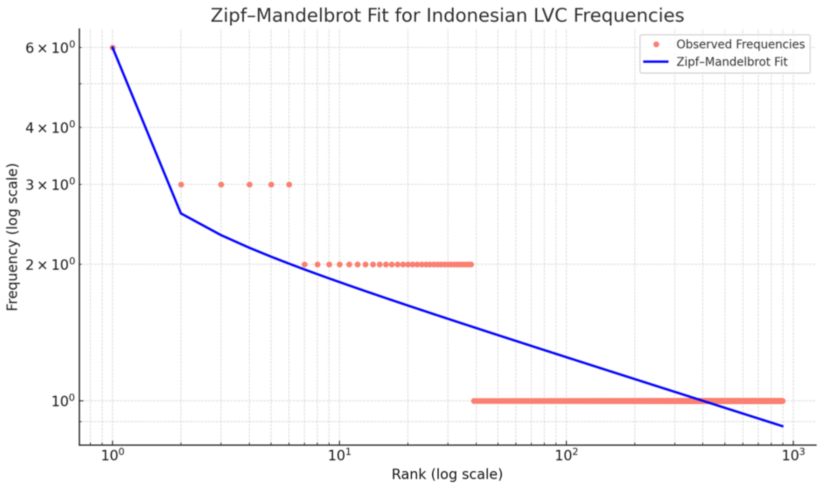


Figure 3.13: Zipf-Mandelbrot fit for LVC frequencies.

Furthermore, the Zipfian analysis reveals a notable discrepancy in the *tail region of the distribution,* where the majority of LVC types occur infrequently. This tail-heavy structure—common in high productive linguistic domains—undermines the linearity assumed by the pure Zipfian form. To address this, as illustrated in Figure 3.13, the Zipf–Mandelbrot law[20] was applied as an extension, incorporating an additional shift parameter to account for non-linear

[20] *See* Altmann and Gerlach (2016), Manin (2009), Piantadosi (2014), and Popescu *et al.* (2010).

head and tail behavior. The model yielded parameters $a \approx 2.59$, $b \approx 0.159$, and $c \approx -0.995$, producing a more accurate fit to the empirical data, particularly in the low-frequency ranks. The low exponent b indicates a relatively flat distribution, consistent with a system where many LVCs share similar, modest usage rates. The slight negative value of c further suggests an adjustment for the *frequency plateau* observed in the lower ranks—deviation that could not be captured by the simples Zipf model. Therefore, the observed deviation in low-frequency LVCs is not merely statistical noise, but rather a manifestation of the *rich tail of the lexicon,* where creative, emergent, or context-specific constructions reside. These constructions are functionally valuable even if infrequent, and their proliferation skews the distribution away from the canonical Zipfian slope. The Zipf–Mandelbrot enables a more faithful modeling of this phenomenon by capturing the *flattened tail* without distorting the overall rank-order hierarchy. This refinement supports a broader view of the lexicon as hierarchically organized, where frequency patterns are shaped by semantic granularity, inherently linguistic context, and morphosyntactic affordances.

3.2 LVCs' distribution in relation to empirical laws of language

This sub-chapter delves into the distributional characteristics of LVC and their potential interplay with established empirical laws of language. While Zipfian Distribution is a fundamental law in the present study, examining the inverse relationship between a LVCs' frequency and its rank, several other laws and measures are used to evaluate frequency and distribution. Specifically, the analysis examines the extent to which the observed patterns in LVC occurrence and frequency align with, or deviate from, the predictions and insights offered by prominent linguistic regularities. This analysis will focus on six key empirical laws: the Menzerath-Altmann Law, positing an inverse relationship between the size of a linguistic construct and the size of its constituent (§3.2.1); Good-Turing Frequency Estimation, a method for estimating the probability of unseen or low-frequency events (§3.2.2); Yule's K, a measure of lexical diversity and concentration (§3.2.3); Heaps' Law, which describes the growth of vocabulary size in relation to text length (§3.2.4); Baayen's morphological productivity framework, evaluating tendencies of the morphological productivity within LVCs (§3.2.5); and Shannon Entropy, quantifying lexical unpredictability and distributional evenness within LVCs dataset (§3.2.6).

3.3.1 The internal structure of LVCs and the Menzerath–Altmann Law

The Menzerath–Altmann Law, a cornerstone of quantitative linguistics, describes an inverse relationship between the size of a linguistic construct and the size of its constituent elements. Often summarized as “*the greater whole, the smaller its parts*,” this principle has been applied across various levels of linguistic organization, from phonology to syntax. In the present study, the law is applied to Indonesian LVCs, which serve as productive multiword expressions comprising a verbal element and a nominal element. This analysis investigates whether longer LVCs, defined at varying linguistic granularities, are composed of less frequent constituents—operationalized here through the inverse of observed frequency (`Inverse_TOTAL`). Four models are developed to reflect different conceptions of *length*: word-based, morpheme-based, syllable-based, and phoneme-based. Each model is fitted using Menzerath–Altmann function $y = a\, xb\, e - cx$, and their comparative fits are assessed using the coefficient of determination (R^2). By visualizing and comparing these models within a unified log-scaled framework, this analysis not only tests the applicability of the law in the context of LVCs but also evaluates which linguistic unit best captures the principle of constituent economy in the Indonesian lexicon.

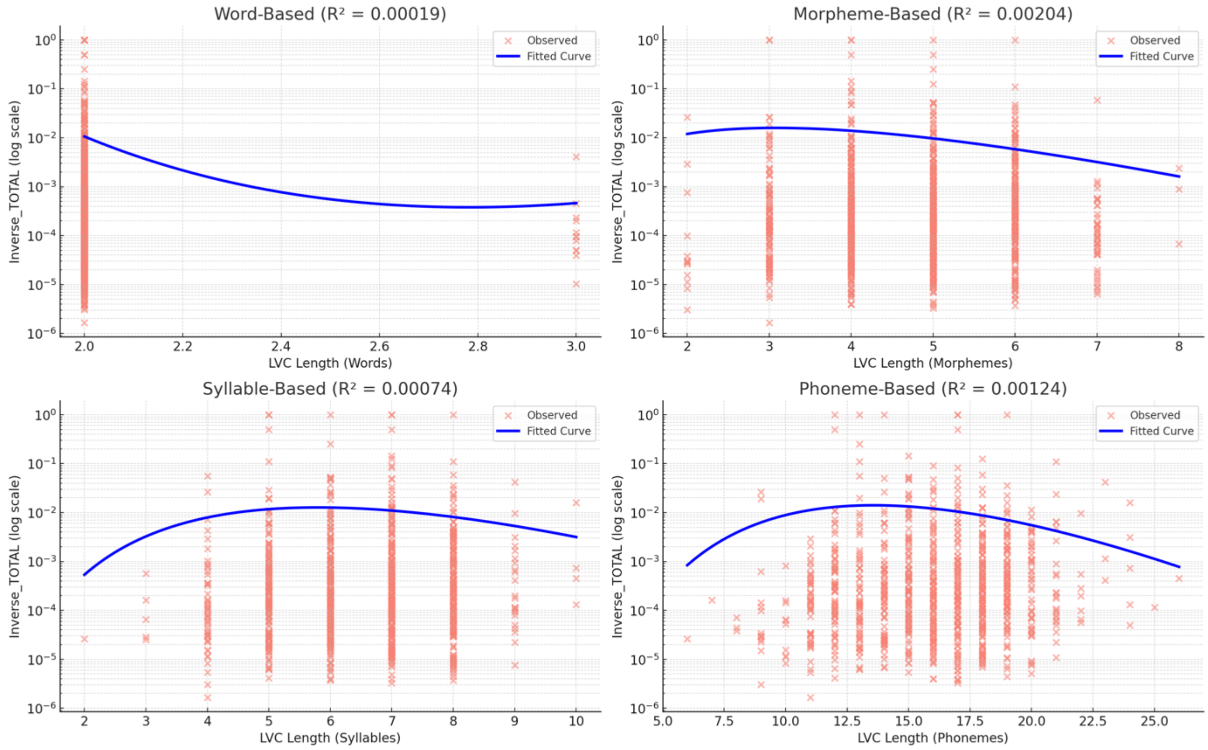


Figure 3.14: Testing the Menzerath-Altman Law using word–, morpheme–, syllable–, and phoneme–based LVC lengths.

First, the word-based model provides a baseline for the analysis by treating each LVC as a sequence of words. In this model (*see* Figure 3.14, upper-left), the number of words as serves as the independent variable, while `Inverse_TOTAL` functions as the dependent measure of constituent weight. The fitted Menzerath–Altmann parameters for this model are a ≈ 2.63 × 10–6, b ≈ 3.56, and c ≈ 0.94, resulting a very low explanatory power with $R^2 \approx 0.00019$. The plotted curve is shallow and barely deviates from a flat trend, indicating that the number of words in an LVCs offers minimal insight into its frequency-based complexity. This result is consistent with prior criticisms that word counts, especially in morphologically rich or agglutinative languages like Indonesian, may not meaningfully reflect linguistic structure or cognitive load. Words in Indonesian often encapsulate multiple layers of information through affixation, making their raw count an imprecise unit of structural measurement. Consequently, the word-based model demonstrates limited alignment with the Menzerath–Altmann hypothesis and highlights the need more granular linguistic representations. The near-zero R^2 further suggests that modeling constituent complexity using surface-level lexical length is inadequate in capturing the deeper distributional regularities of LVCs.

Second, among four models evaluated, the morpheme-based approach yields the strongest alignment with the Menzerath–Altmann Law (*see* Figure 3.14, upper-right).[21] Here, LVC length is defined as the total number of morphemes per construction, accounting for prefixes, suffixes, and circumfixes based on Indonesian morphological rules. The model estimates are a ≈ 0.01029, b ≈ 3.54, and c ≈ 1.15, with corresponding $R^2 \approx 0.00204$—the highest among all configurations tested. While still modest by conventional standards, this value reflects a tenfold improvement over the word-based model. The curve fitted to the morpheme-based data reveals a clearer inverse trajectory, indicating that longer constructions, in terms of morphological complexity, tend to occur less frequently. This aligns well with cognitive-linguistic interpretations that posit a processing cost associated with morphologically rich expressions, which may be offset by lower usage frequency. The morpheme-based measure captures the functional layers embedded in each LVC, such as agency, aspect, or valency, offering a structurally sensitive and theoretically robust unit of analysis. This result validates the choice to model LVCs through a morphological lens and reinforces the theoretical premises

[21] Altmann's (1967) notable quantitative-linguistics work entitled "The structure of Indonesian morphemes" became one of the basic morphemic parameters used in this study. Considering that Indonesian has developed considerably since the year the study was conducted, the present study made adjustments by adding updates from several recent literatures.

that linguistic economy operates not at the level of surface tokens but within the architecture of meaningful subunits.

Third, the syllable-based model operationalized LVC length as the number of syllables per construction, using a simplified Indonesian syllabification rule (e.g., V, C, CV, and CVC) (*see* Figure 3.14, bottom-left). The resulting model exhibits a better fit than the word-based version but remains weaker than its morpheme-based counterpart. The estimated parameters are a ≈ 3.60 × 10–5, b ≈ 7.75, and c ≈ 1.34, yielding a coefficient of determination $R^2 \approx 0.00074$. While this reflects a nearly fourfold improvement over the word-based model, the low R2 suggests that syllabic length only partially captures the structural patterns predicted by the Menzerath–Altmann Law. The fitted curve displays a discernible downward trend on the log-scaled plot, indicating some support for the law's inverse hypothesis. However, syllables, as phonological units, may vary informational density and articulatory complexity, making them a less consistent proxy for constituent structure. Despite these limitations, the model retains some interpretative value, especially in relation to prosodic or fluency-based theories of language production. It provides evidence that even at the level of syllabic realization, there exists a weak tendency for longer expressions to be composed of simpler phonological elements.

Fourth, the phoneme-based model explores the most granular level of linguistic length by counting phoneme-like units within each LVC (*see* Figure 3.14, bottom-right). Using a tailored inventory of Indonesian phonemes[22], including diphthongs and consonant clusters (e.g., *ng, sy, kl*), this model assesses whether segmental length correlates inversely with frequency. The estimated parameters are a ≈ 3.35 × 10–10, b ≈ 10.91, and c ≈ 0.80, producing an $R^2 \approx 0.00124$. This makes it the second-best fitting model after the morpheme-based approach. The log-scaled plot reveals a smoother and more defined curve compared to the syllable- and word-based versions, suggesting that segmental density may serve as a meaningful, though indirect, reflection of constructional complexity. While phonemes do not map directly onto semantic structure, their accumulation within longer LVCs appears to correlate with a general decrease in frequency, consistent with the Menzerath–Altmann expectation. This model offers a fine-grained lens through which to examine phonological realization, especially in studies that bridge structural linguistics with speech production or cognitive processing. Although the

[22] Except for diphthongs and clusters, the phoneme list used in this analysis was also used by Altmann (1966) in his notable study entitled "Binomial index of euphony for Indonesian poetry."

phoneme-based model does not outperform the morpheme-based configuration, it contributes to a layered understanding of linguistic economy at multiple representational levels.

The comparative of the Menzerath–Altmann Law across word-, morpheme-, syllable-, and phoneme-based measures offers nuanced insight into the structural dynamics of Indonesian LVCs. Empirically, the morpheme-based model emerges as the most robust representation, providing the clearest fit to the inverse law with $R^2 \approx 0.00204$. The morpheme-based model follows closely, suggesting that both morpho-semantic and phonological considerations are essential to understanding constituent complexity. The syllable-based model, while partially aligned, underperforms relative to morphemes and phonemes, and the word-based model offers the weakest support. These findings validate the theoretical premise that linguistic economy is best observed at structural—not surface—levels of representation. From a methodological perspective, the analysis also highlights the importance of granularity: finer-grained unites (morphemes and phonemes) capture patterns that coarser units (words and syllables) tend to obscure. The implications are significant for both corpus linguistics and construction grammar, suggesting that future modeling efforts should prioritize morpho-structural metrics when testing complexity laws. In short, this multi-level evaluation of the Menzerath–Altmann Law in Indonesian LVCs underscores the layered nature of linguistic organization and offers a framework for extending structural economy analyses across languages and constructions.

Notably, Altmann's early empirical studies in 1966–1967 focused on the Indonesian language, offering one of the earliest applications of quantitative linguistic methods to non-European morpho-phonological systems. His analysis, based on dictionary-level lexical structure, served as a benchmark for understanding morpheme complexity, phoneme inventory, and structural class-group correlations in Indonesian. In contrast to such lexical forms, this analysis investigates the naturally-occurring LVC structure in modern Indonesian by analyzing empirical data from a large corpus of verbal-nominal constructions. Within the scope of the sub-chapter, the internal structure of LVCs and the Menzerath–Altmann Law, a specific framework is constructed to assess how contemporary usage diverges from lexicographic norms. This contrast sheds light not only on the validity of the Menzerath–Altmann principle in productive constructions but also on how real-world usage conditions reshape the distribution and complexity of lexical units.

Table 3.12: Altmann's (1967) dictionary-based lexical model of Indonesian and empirical metrics from contemporary Indonesian LVCs.

Metric	**Dictionary-level Lexical Structure** (Altmann, 1967: 36)	**Naturally-occurring LVCs' Structure**	**Note**
Number of phonemes in inventory	29	29	Shared phonological baseline (Indonesian)
Number of morpheme types	143	935	Broader lexical variation in LVCs
Max. phonemes in a morpheme (proxy)	12	26	More phonologically dense items in LVCs
Max. syllables in a morpheme (proxy)	3	10	Indicates compound or elaborated forms
Avg. type length in phonemes	6.9371	15.9055	LVCs are more than twice as long as dictionaries entry
Avg. type length in syllables	2.9510	6.4745	Same doubling pattern in syllabic complexity
Total number of morphemes	11,006	4,488	Corpus scope is narrower but more dynamic
Avg. morphemes length in phonemes	5.3399	3.3385	Shorter internal elements due to high-frequency affixes
Avg. morphemes length in syllables	2.1646	1.3590	Reflects morphological compression in usage
Mean group range W_G	4.60	134.5714	LVCs cluster tightly into shared length patterns
Mean class range W_C	1.33	5.0000	LVCs structurally longer than standard dictionary forms
Regression of G on C	G = 0.3399 + 0.3764C	G = 174.2143 – 7.9286C	Reverse trend: group size decreases with complexity in LVCs
R^2 (G on C)	*not reported*	0.0206	Weak predictability
Regression of C on G	C = 2.1860 + 1.6100G	C = 5.3495 – 0.0026G	Altmann shows class rising with group; LVCs show flattening
R^2 (C on G)	*not reported*	0.0206	Suggests structural flexibility in LVCs
Pearson's Correlation (*r*)	0.7784	≈ 0.1436	Strong dictionary-based correlation vs. weak empirical dependency

As presented in Table 3.12, the comparative analysis reveals striking contrasts between Altmann's dictionary-derived values and the observed characteristics of naturally-occurring LVCs. Altmann documented an average morpheme length of 5.3399 phonemes and 2.1646 syllables, based on 11,006 morphemes from standard lexical entries. In contrast, the LVC dataset exhibit significantly shorter average morphemes—3.3385 phonemes and 1.3590 syllables—despite being drawn from more complex verbal constructions. This suggests a functional compression within LVCs, potentially driven by high-frequency affixes such as {*me-*}, {*di-*}, and {*-kan*}, which served to encode syntactic relations without increasing phonological burden. Interestingly, while LVCs have shorter morphemes, they are structurally longer units overall. The average LVC length is 15.9055 and 6.4745 syllables, compared to Altmann's average morpheme length of 6.9371 phonemes and 2.9510 syllables. These findings affirm that LVCs are semantically and syntactically extended constructions, characterized by more phonological material, despite relying on smaller morphological building blocks. This phenomenon highlights how naturally-occurring LVC structure prioritize expressivity and grammatical function over compactness observed in dictionary-level lexical structure.

Another axis of comparison involves the distributional organization of morpheme lengths and their frequency-based groupings. Altmann's analysis produced a mean group range (WG) of 4.60, indicating a relatively modest number of morphemes clustering around particular classes. In contrast, the LVC data demonstrates a dramatically higher WG = 134.5714, reflecting a high degree of length-based concentration within a narrower set of structural patterns. Similarly, the mean class range (WC) in Altmann's findings was 1.33, compared to 5.0000 in the LVC dataset, signalling that naturally-occurring LVCs span a broader morphological length spectrum than standard lexical entries. This shift points to a critical functional divergence: while dictionary-defined morphemes are curated for typological breadth, LVCs reflect production-level efficiency, where certain length template are recycled across constructions. These findings underscore how LVCs operationalized linguistic economy not by compressing structural units, but by densely populating a specific morpho-phonological window that is both productive and cognitively tractable.

Perhaps the most telling contrast lies in the statistical relationships between class size and group frequency. Altmann identified a strong positive correlation ($r = 0.7784$) between group and class values, suggesting a robust structural regularity within the dictionary-based lexical inventory. In the LVC data, however, this correlation drops dramatically to $r \approx 0.1436$, with both regression models yielding minimal explanatory power (R2 = 0.0206 for both G on C and C on G). The equations derived from the LVC dataset, $G = 174.2143 - 7.9286C$ and $C = 5.3495 - 0.0026G$, exhibit inverse trends to Altmann's original formulas, indicating that longer constructions tend to occur less frequently—a pattern that aligns with Menzerath–Altmann principle but diverges from lexicographic norms. This suggests that naturally-occurring LVC structures is less tightly governed by morpho-statistical regularity and more influenced by semantic utility and discourse frequency. In other words, the predictive coherence observed in curated dictionaries is replaced by probabilistic fluidity in actual usage.

Together, these metrics illustrate a compelling shift from dictionary-level lexical structure to naturally-occurring LVC structure. The comparison affirms that while dictionary entries optimize for representational clarity, real-world constructions such as LVCs embody adaptive strategies for communication. The higher phonological length and lower morphemic density in LVCs reveal a tendency toward syntactic packaging and semantic extension, whereby complex ideas are expressed through relatively simple, affixed forms arranged in predictable templates. At the same time, this weak statistical regularity in group-class relations and the higher concentration of frequency around mid-length LVCs point to a usage-based grammar—one shaped more by discourse function and collocational tendency than by structural equilibrium. Thus, while Altmann's model provides an essential reference point for understanding lexical architecture, the data presented here demonstrate that constructions like LVCs evolve their own distributional logic, one that continues to conform broadly to the Menzerath–Altmann Law, but in ways that are shaped by communicative need rather than formal paradigms.

3.3.2 An application of Good-Turing Frequency Estimation

The Good–Turing Frequency Estimation[23] method offers a statistically principled approach to addressing the problem of unseen or rare events in linguistic data. This method adjusts the raw frequency counts or observed items to better approximate the true underlying distribution of a population, especially in domains where the sample may not fully capture the system's generative possibilities. In linguistic applications, such as corpus-based analysis of multiword constructions or lexical items, the Good–Turing method serves as a robust tool for estimating the likelihood of unobserved but theoretically plausible expressions. Rather than relying solely on the empirical frequency of each item, the technique models how often items of a given frequency (e.g., those occurring once, twice, etc.) appear and uses that meta-distribution to redistribute probability mass. This particularly important in context involving highly skewed frequency distributions, such as LVCs, where a small number of collocations dominate usage while the long tail consists of numerous infrequent or unattested items. The Good–Turing method not only smooths frequency estimates but also allows for the quantification of unseen types—those constructions which are not present in the corpus but are presumed to exist within the larger linguistic system. Given the productivity and compositionality of LVCs in Indonesian, applying the Good–Turing framework provides a perspective on the structural richness of the lexicon and the completeness of the current dataset.

[23] *See* Gale and Sampson (1995), Good (1953), and Hwang *et al.* (2015).

The initial implementation of the Good–Turing estimation involved analyzing all observed LVCs in dataset by first generating a frequency-of-frequencies table. This table indexed how many unique LVCs occurred exactly once, twice, and so on, forming the basis for adjusted frequency estimates *r**. For example, LVCs with a raw frequency of 1 (hapax legomena) were assigned an adjusted frequency of approximately 0.67, while those occurring twice were adjusted to 1.5, and those occurring four times were raised to 5.0. These adjustments reflect a redistribution of frequency weight from high-frequency items to lower-frequency ones, in line with the assumption that rare items are underrepresented due to corpus limitations (*see* Figure 3.15). The fitted model allows for a more probabilistically sound estimation of each LVC's presence in the language, even when data sparsely is an issue. Notably, the model preserves the empirical distribution while adjusting for bias in the tails, which is crucial for fairer representations in downstream analyses such as clustering or linguistic profiling. In highly productive systems like LVCs, where new combinations can emerge through compositional means, such smoothing helps us to latent importance of infrequent ones. This step also lays the groundwork for estimating unseen events, a core strength of Good–Turing method.

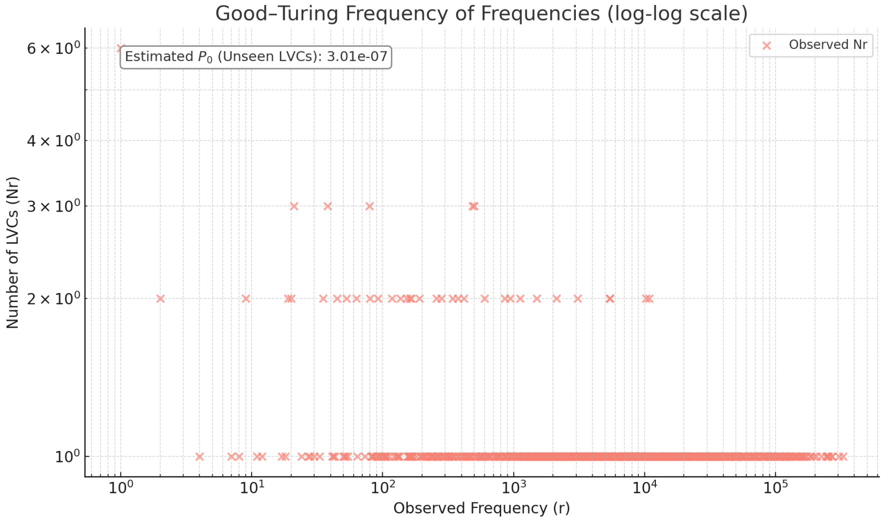


Figure 3.15: Good-Turing frequency-of-frequency plot in log-log scale, which is visualizes how many LVCs occur with each observed frequency (*r*).

To sharpen the analytical focus, a subset of rare LVCs—those with observed frequencies of ten or fewer—was isolated for closer inspection. This segment of the data offers a window into the tail of the distribution where linguistic creativity and structural innovation are often most visible. Within this subset, the Good–Turing method reveals considerable upward adjustments to the frequency of rare constructions. For instance, LVCs with a raw count of 8 were adjusted to frequency of 18.0, and those with 7 occurrences were adjusted to 8.0. These adjustments suggest that, although such constructions are rare in the observed dataset, their likelihood of recurring is higher than their raw frequencies might imply. This is consistent with the notion that rare items are often under-sampled in corpus data and may still hold relevance in the mental lexicon or broader speech community. By applying smoothing selectively to this low-frequency region, we can gain a more realistic view of the distributional potential of infrequent LVCs without inflating the significance of highly frequent items. Importantly, these adjustments serve not only as technical corrections but also as theoretical indicators for underlying linguistic patterns. They provide a principled way of acknowledging that some LVCs may appear rare only due to the finite nature of the corpus and not because they are marginal in actual usage.

One of the implications of the Good–Turing method is ability to estimate the probability mass associated with unseen events—in this case, LVCs that presumably do not appear in the dataset at all. Based on the presence of six hapax legomena (LVCs observed exactly once), the probability of encountering an unseen LVCs was estimated to be approximately P0 ≈ 3.01×10-7. This is an extremely small probability, suggesting that the current dataset is highly comprehensive and captures nearly the entire productive space of observable LVCs within the sample. Using this probability estimate in conjunction with the number of observed LVC types, the projected number of unseen LVCs in the language was calculated to be approximately 0.00028. This minuscule figure indicates that the likelihood of encountering a novel LVC not present in the dataset is exceedingly low—an important finding for evaluating corpus completeness and model generalizability. From practical standpoint, this result affirms that the observed LVC inventory saturates the data distribution and that further data collection is unlikely to yield a substantially different set of constructions. Theoretically, it supports the notion that the majority of structurally plausible LVCs in the language are already attested, thereby strengthening the readability of distributional analyses based on this corpus. In sum, the application of Good–Turing estimation has provided both a corrective lens of interpreting rare constructions and a quantitative basis for asserting the sufficiency of the dataset in representing the Indonesian LVC system.

3.3.3 Notable insights from Yule's K

Lexical concentration, as captured by Yule's K statistic[24], offers a lens through which to evaluate the diachronic variation of multiword constructions within a corpus. Yule's K is a measure of lexical richness based on the frequency of frequencies of words in a text. The formula is:

$$K = 10^4 \text{ x } \sum_{i=1}^{V} i^2 V_i - N \qquad (3.3)$$

Where:
- V is the number of types
- V_i is the total number of types that occur i times
- N is the total number of tokens.

In this study, Yules' K is employed to trace changes in the distributional density of LVCs across four Indonesian corpora: SLIC (2010), IWC (2012), ILCC (2013), and IDC (2020). This diachronic perspective allows us to observe whether LVC usage becomes increasingly diverse or repetitive over time, revealing patterns in linguistic innovation, formulaicity, and register-driven regularity. Given that Yules' K increases as repetition intensifies, a rising trend in K-values would indicate lexical entrenchment, whereas a decline would reflect diversification and structural expansion of LVCs (*see* Figure 3.16). The present analysis reveals substantial fluctuation in Yules' K across the decade, suggesting non-linear changes in lexical density. These variations can be tied to shifts in corpus composition, genre distribution, or broader sociolinguistic dynamics such as the digitization of discourse and evolving syntactic preferences. Crucially, this approach treats lexical concentration not as a static feature of the language but as a historically contingent outcome of communicative, stylistic, and structural pressures acting on collocational behavior.

The computed Yule's K values offer a quantifiable basis for comparing lexical repetition across the four corpora. The earliest one, SLIC (2010), exhibits the lowest concentration with a Yule's K of 51.06, followed closely by IWC (2012) at 51.92. these values suggest a period of high lexical diversity, characterized by a relatively balance distribution of LVC types and a lower recurrence of fixed expressions. However, this pattern shifts markedly in ILCC (2013), where the Yules' K rises sharply to 65.01—the highest among all four corpora. This peak in

[24] *See* Jarvis (2013, 2017), Jarvis and Hashimoto (2021), Koizumi and In'nami (2012) and Kyle *et al.* (2021).

lexical concentration suggests a temporary phase of lexical saturation of formulaic reliance, possibly driven by the nature of the ILCC corpus content (e.g., formal documents, educational materials) or a stylistic preference for repeated constructions. In contrast, the IDC (2020) corpus shows a modest reduction in concentration, with a Yule's K of 54.59. This return to lower repetition levels may reflect the influence of more recent language usage norms, including informalization, digital genre discourse. The diachronic curve, as visualized in the enhanced plot, captures a non-monotonic trend: lexical diversity increases from 2010 to 2020, peaks in formulaicity in 2013, and then re-diversifies by 2020.

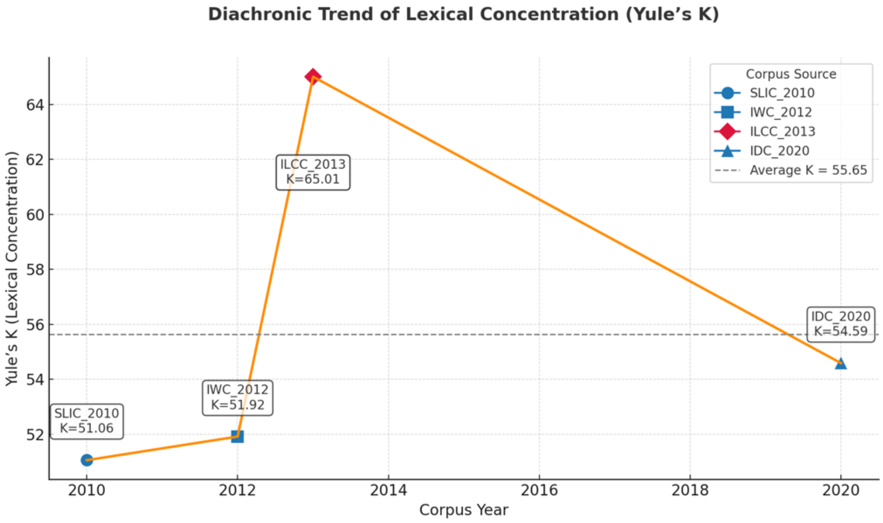


Figure 3.16: Diachronic trend of lexical concentration based on Yule's K.

Furthermore, to deepen the diachronic investigation of LVCs dynamics in Indonesian, this study incorporates Kullback-Leibler (KL) Divergences[25] as a probabilistic measure of distributional change. While prior metrics such as Yule's K capture lexical concentration, KL Divergence uniquely quantifies how one LVC distribution diverges from another. It offers a directional, asymmetric comparison—answering how surprising or information-costly one corpus would be if interpreted through the frequency expectations of another. Because raw KL Divergence is sensitive to zero probabilities, a *laplace smoothing procedure* was applied to stabilize the comparison across all corpus years: SLIC_2010, IWC_2012, ILCC_2013, and IDC_2020. The resulting heatmap visualizes pairwise divergences, unveiling not only *lexical shifts across time,* but also the *lexical stabilization between specific periods.* This approach

[25] *See* other works in Aradilla *et al.* (2008), Imseng *et al.* (2012), Prokhorov *et al.* (2019) and Wu *et al.*, (2025).

directly complements findings from Yule's K, offering a second line of evidence to tract shifts in lexical concentration and the emergence or dissolution of formulaic usage over time.

The KL Divergence metric identifies SLIC_2010 as a lexical outlier within the diachronic trajectory of LVC development. Its divergence from other corpora is consistently the highest: D(SLIC_2010|IDC_2020) = 2.0419, D(SLIC_2010|ILCC_2013) = 2.0263, and D(SLIC_2010|IWC_2020) = 1.8071. These values indicate that interpreting later corpora (Q) through the lens of 2010's frequency expectations (P) would incur high informational cost—signaling substantial shifts in LVC usage. These findings align with SLIC_2010's Yule's K of 51.06, indicating relatively low concentration, but the KL values suggest that the diversity in 2010 was have featured not only wider range of LVCs, but a compositionally unique distribution profile. This uniqueness might stem from genre-specific features, sociolinguistic conventions of the early digital era, or reduced lexical entrenchment. The divergence values reveal that *lexical change over the decade is not merely additive or incremental,* but directional and potentially genre-driven. The heatmap as in Figure 3.17 distinctly shows that 2010 does not align closely with any subsequent year, making it a lexical pivot point in the evolution of Indonesian LVCs.

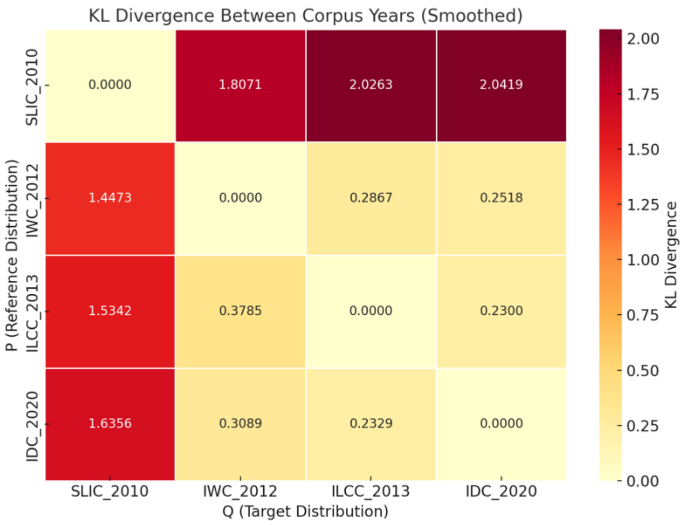


Figure 3.17: KL Divergence between corpus years.

While SLIC_2010 stands apart, the KL Divergence analysis shows increasing convergence among IWC_2012, ILCC_2013, and IDC_2020, suggesting a gradual stabilization of LVC usage. The *lowest divergence* found between ILCC_2013 and IDC_2020, with D(ILCC_2013|IDC_2020) = 0.2300, and vice versa at D(IDC_2020|ILCC_2013) = 0.2329. Similarly, IWC_2012 shows low divergence from IDC_2020 (D=0.2518) and ILCC_2013

(D=0.2867). These values indicate that over time, the LVC distributions in Indonesian corpora have begun to stabilize, possibly due to shared educational, journalistic, or institutional norms. Notably, this trend aligns with Yule's K, which peaked in 2013 before modestly decreasing in 2020. The convergence in KL values suggests that while individual LVC types may rise or fall, the *overall shape of the distribution has normalized,* reflecting communicative consolidation. These findings point toward the crystallization of certain LVC patterns as stylistic conventions within written Indonesian, offering empirical support for the notion *formulaic entrenchment* following a period of lexical expansion.

The application of KL Divergence adds theoretical depth and methodological precision to the study of lexical change in LVC usage. Unlike frequency-based metrics, KL Divergence directly models distributional drift, capturing how LVCs patterns diverge across time not merely in terms of variety, but in how those varieties are weighted. The asymmetry of KL also allows for directional insights—for example, the fact that D(SLIC_2010|IDC_2020) = 2.0419 while D(IDC_2020|SLIC_2010) = 1.6356 suggest that 2020 retains elements of 2010, but 2010's patterns are increasingly unrecognizable through the 2020 lens. This directional tension reveals a trajectory of lexical reorganization and possible functional pruning in the Indonesian lexicon. The incorporation of KL Divergence thus supports a multi-layered model of diachronic lexical concentration: while Yule's K charts compactness, KL captures dynamic misalignment, quantifying how far the lexicon has traveled. Thus, in the context of this analysis, KL Divergence reinforces the claim that Indonesian LVCs are not only structurally diverse but historically dynamic—subject to quantifiable shifts in usage intensity and distributional balance.

3.3.4 An examination through Heaps' Law

Heaps' Law (or Herdan's Law)[26] describes the relationship between the size of a corpus (number of tokens) and the number of unique words (type). It suggests that as a corpus grows, the vocabulary size increases, but at a decreasing rate. The formula is often expressed as:

$$V = Kn^{\beta} \tag{3.2}$$

[26] For further readings, refer to Egghe (2007), Egghe and Rousseau (2003), Ross and Herdan (1960), Serra-Peralta *et al.* (2021) and Zhang *et al.* (2016).

Where:

- V is the number of types
- n is the total number of tokens
- K and β are constants that depend on the specific corpus and language. β is typically between 0 and 1.

Given the data structure of the present study, an adaptation of Heaps' Law can be employed to investigate the relationship between the number of distinct lexical varieties (LVC types) and their corresponding cumulative frequencies, wherein each row is construed as a discreate *type* representing a specific lexical variety of LVC and the *total column* denotes its frequency of occurrence. Consequently, several analytical procedures become feasible: the summation of the *total column* yields the aggregate number of tokens within the dataset, while the total count of rows directly furnishes the overall number of distinct lexical varieties (LVC types). Although a direct application of Heaps' Law in its conventional cumulative formulation is not strictly feasible, an exploration of the relationship between the rank of a lexical variety (as indicated by the RANK column) and its frequency (TOTAL column) can be undertaken to ascertain the presence of a power-law-like distribution, a phenomenon intrinsically related to Heaps' Law.

Accordingly, the Figure 3.18 offers a comparative perspective on the cumulative distribution of observed LVCs against the theoretical trajectory predicted by Heaps' Law. The empirical data, represented by discrete blue markers, illustrates the accumulated number of distinct LVC types as a function of the cumulative total frequency within the dataset. Juxtaposed against this empirical observation is the orange line, delineating the expected growth in lexical variety as posited by Heaps' Law, parameterized by the estimated values of K (7.387577325128133e-16) and β (2.44776617974269). A visual inspection reveals a marked divergence between the observed LVC accumulation and the anticipated growth curve dictated by the Heapsian model. In typical linguistic corpora, Heaps' Law manifests as a decelerating increase in the number of unique lexical items with increasing corpus size, characterized by an initial rapid expansion followed by a gradual plateauing. However, the depicted empirical data exhibits a distinct pattern, suggesting a fundamentally different relationship between frequency accumulation and the emergence of novel LVC forms within this specific dataset. This initial observation necessitates a more granular examination of the estimated parameters and the underlying characteristics of the data that might account for this deviation from the conventional Heapsian growth pattern.

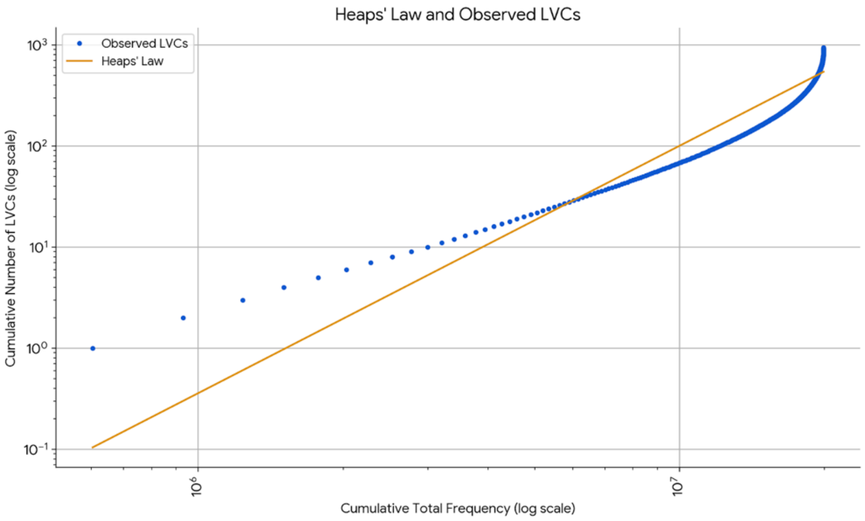


Figure 3.18: Atypical growth of LVC varieties compared to Heaps' Law.

The estimated parameters for Heaps' Law, derived from the present dataset, provide crucial quantitative insights into the observed deviation. The *β* exponent, with a notably elevated of 2.44776617974269, stands in stark contrast to the typical range of $0 < \beta < 1$ reported in studies of natural language corpora. This unusually high *β* value signifies an exceptionality rapid rate at which new LVC types are introduced as the cumulative total frequency increases. In essence, for each increment in the overall frequency count, the number of distinct LVCs grows at a substantially accelerated pace compared to what is generally observed in the vocabulary growth of evolving texts. Conversely, the estimated value of *K,* a scaling factor, is exceedingly small (7.387577325128133e-16). This diminutive *K* value is not necessarily an intrinsic property of the underlying linguistic phenomena but rather a mathematical consequence of the scaling inherent in the data and the exceptionally high *β* exponent. Given the multiplicative relationship between *K* and the frequency raised to the power of *β* in the Heaps' Law equation ($V = KN^{\beta}$, where *V* is vocabulary sized an *N* is corpus size), a large *β* necessitates a correspondingly small *K* to fit the overall scale of the data. Therefore, the atypical parameter values collectively underscore the significant departure of the observed LVC distribution from the patterns typically captured by Heaps' Law in standard corpus linguistic analyses.

Several potential factors could account for the observed discrepancy between the empirical LVC distribution and the predictions of Heaps' Law. Firstly, the inherent of the data, comprising a pre-defined set of lexical varieties and their associated frequencies, fundamentally differs from the dynamic growth of a traditional text corpus. Heaps' Law is conventionally applied to analyze the vocabulary expansion as a text unfolds and new words are encountered.

The current dataset, however, represents a static collection where the lexical units (LVCs in this case) are already established, and the analysis focuses on their frequency of occurrence. Consequently, the application of Heaps' Law here is an adaptation, and deviations are not expected. Secondly, the specific methodologies employed in defining and preprocessing the LVCs would exert a considerable influence on the resulting counts and distributions. Variations in lemmatization (applicable to LVCs as multiword units), and the criteria used to identify and categorize LVCs could lead to unique distributional characteristics. Finally, the intrinsic linguistic properties of the language under investigation might contribute to the observed patterns.

3.3.5 An examination through Baayen's framework

To complement the structural and frequency-based analyses of LVCs, this study incorporates Baayen's morphological productivity framework[27]. This approach enables an evaluation of linguistic creativity and lexical innovation by focusing on the distribution of hapax legomena—types that occur only once in the dataset. Three key metrics are utilized: realized productivity ($P = V/N$), expanding productivity ($S = V_1/V$), and potential productivity ($V = V_1/N$, synonymous with P in this case). The analysis is applied to both LVC clusters and diachronic corpus segments to assess how morphological richness is distributed across linguistic groupings and over time. The use of log-scaled heatmaps enhances the interpretability of these data, as it enables subtle differences in productivity values—often spanning multiple orders of magnitude—to be effectively visualized and compared. This method is particularly useful in the present study, where most productivity measures are extremely low but still theoretically significant. By translating minimal numerical values into perceptible visual contrast, the log transformation underscores emergent patterns of morphological expansion or stagnation in a statistically robust way.

The heatmap visualization reveals stark contrast in morphological productivity across the three LVC clusters. Notably, Cluster 1 stands out as the only group exhibiting non-zero values across all three productivity dimensions, with $P \approx 5.79 \times 10^{-7}$, $S = 0.0071$, and $V \approx 5.79 \times 10^{-7}$. These values, while low in absolute terms, are significant within the domain of morphological analysis, especially when compared to Cluster 2 and Cluster 3, both of which show zero

[27] *See* Baayen (1989, 1992, 2009), Baayen and Lieber (1991).

productivity on all dimensions. The expanding productivity (S) in Cluster 1 is especially noteworthy, indicating that approximately 0.71% of the cluster's LVC types occur only once—suggesting latent morphological creativity. This aligns with prior findings from Yule's K and K-means Clustering, which also identified Cluster 1 at the most lexically diverse and least formulaic grouping. In contrast, Clusters 2 and 3 are likely composed fixed, repetitive constructions with entrenched morphological forms that do not generate novel expressions. The sharp visual contrast in the heatmap (*see* Figure 3.19), with deep coloration for Cluster 1 and near absence of color for the others, makes this differentiation immediately salient. This finding supports the hypothesis that LVC usage in Indonesian is not morphologically uniform but instead clustered around distinct modes of productivity.

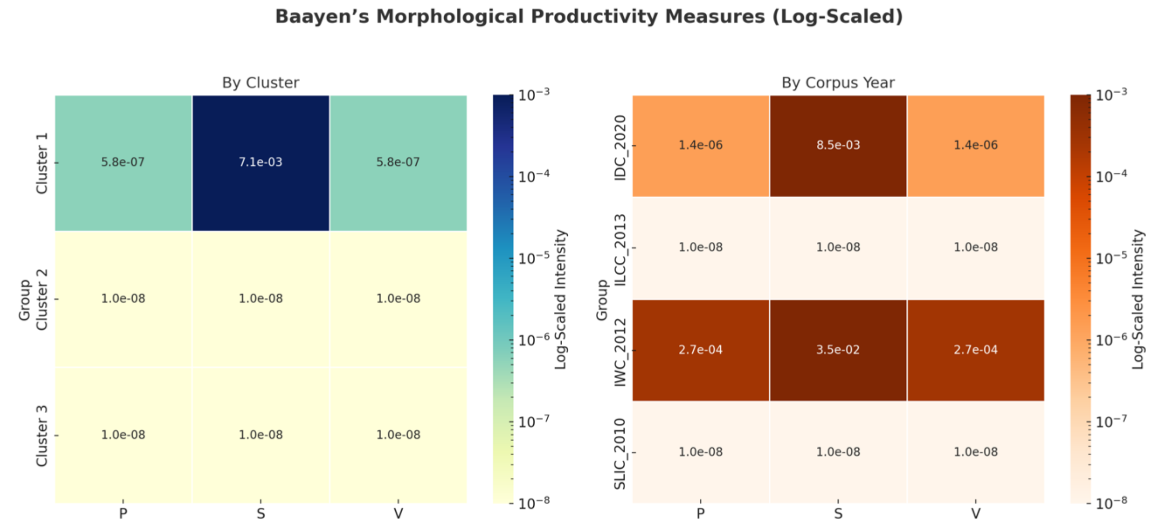


Figure 3.19: Baayen's morphological productivity measure.

The corpus-based heatmap further demonstrates how morphological productivity has varied over time. IWC_2012 emerges as the most morphologically innovative corpus segment, with $P\approx2.74\times10^{-4}$, S=0.0350, and $V\approx2.74\times10^{-4}$. These values indicate that 3.5% of its LVC types are hapax legomena, a clear marker of linguistic expansion and novel lexical information. This peak in productivity may reflect shifts in register or discursive style in 2012—possible tied to digital communication, media, or educational texts that foster lexical experimentation. IDC_2020, though far less productive than IWC_2012, still shows non-zero values ($P\approx1.45\times10^{-8}$, S=0.0085, $V\approx1.45\times10^{-8}$), suggesting a re-emergence of morphological innovation in the most recent data. Meanwhile, SLIC_2010 and ILCC_2013 exhibit no morphological productivity, with all values at zero. This finding correlates with the higher lexical concentration identified in ILCC_2013 (Yule's K≈65.01), which points to formulaic or standardized usage patterns. The log-scaled heatmap visually accentuates these temporal distinctions, with IWC_2012's warm orange tones contrasting sharply against illustrate that morphological innovation in Indonesian LVCs is not temporally static but instead fluctuates in response to socio-discursive dynamics.

These findings collectively underscore the importance of viewing morphological productivity as a dynamic and differentiated property of the lexicon—one that varies across both structural clusters and temporal periods. The integration of Baayen's measures into the analysis reveals dimensions of linguistic creativity that would otherwise remain obscure under frequency-based or structural models alone. The extremely low—but not negligible—values of realized and expanding productivity in Cluster 1 and IWC_2012 highlight zones of morphological innovation that are statistically subtle but theoretically significant. These zones serve as incubators for lexical change, suggesting that even rare constructions contribute meaningfully to the evolution of multiword expressions. The application of log-scaled heatmaps was instrumental in unveiling these distinctions, allowing the nuanced contribution of low-frequency and hapax-based constructions to be rendered visible. This methodological choice complements earlier analyses grounded in the Menzerath–Altmann law, Yule's K, and Zipf–Mandelbrot modeling, offering a holistic view of how LVCs in Indonesian operate across both structure and time.

3.3.6 An examination through Shannon Entropy

Shannon Entropy offers a foundational information-theoretic metric for quantifying lexical unpredictability and distributional evenness within language data.[28] Unlike frequency of productivity measures that focus on repetition or morphological innovation, entropy captures how evenly distributed LVCs are across a given dataset. High entropy indicates a lexicon in which LVC usage is diverse and relatively balanced in frequency, while low entropy reflects skewed distributions dominated by a small number of recurrent items. In the context of Indonesian LVCs, this metric serves as a critical complement to Yule's K and Baayen's P, offering an additional lens through which lexical diversity and communicative richness can be assessed. Entropy values are particularly valuable when comparing linguistic patterns across structural clusters and diachronic corpora, allowing for precise quantification of lexical

[28] For further readings, refer to Arora *et al.* (2022), Bentz *et al.* (2017), Diessel (2017), Pilgrim *et al.* (2024) and Shannon (1948, 1951).

evenness in different communicative contexts. By applying Shannon Entropy at the cluster and corpus-year levels, this analysis seeks to uncover the underlying probabilistic structure of LVC usage, thereby enhancing our understanding of how formulaicity and flexibility vary across the Indonesian lexicon.

The entropy analysis at the cluster levels reveals substantial variation in the distributional structure of LVC usage. Cluster 1, with a Shannon Entropy of 8.47, emerges as the most information-rich and lexically diverse group. This high entropy value suggests that LVCs within this cluster are used with a relatively balanced frequency. In contrast, Cluster 2 registers the lowest entropy score of 4.72, reflecting a heavy reliance on a small set of highly frequent LVCs—an indicator of linguistic formulaicity and predictability. Cluster 3, with an entropy value of 5.62, occupies an intermediate position, implying a mixture of concentrated and distributed usage patterns. These entropy levels align closely with findings from previous metrics: Cluster 1 was earlier shown to have the lowest Yule's K (≈40.08) and the only cluster with non-zero Baayen productivity values, all of which point to high lexical diversity and creativity. The entropy heatmap further highlights this distributional divergence, with Cluster 1 displaying deep coloration indicative of information density, while Clusters 2 and 3 appear paler, visually reinforcing their restricted lexical behavior (*see* Figure 3.20). This confirms that the structural segmentation of LVCs corresponds to functionally distinct regimes of lexical usage.

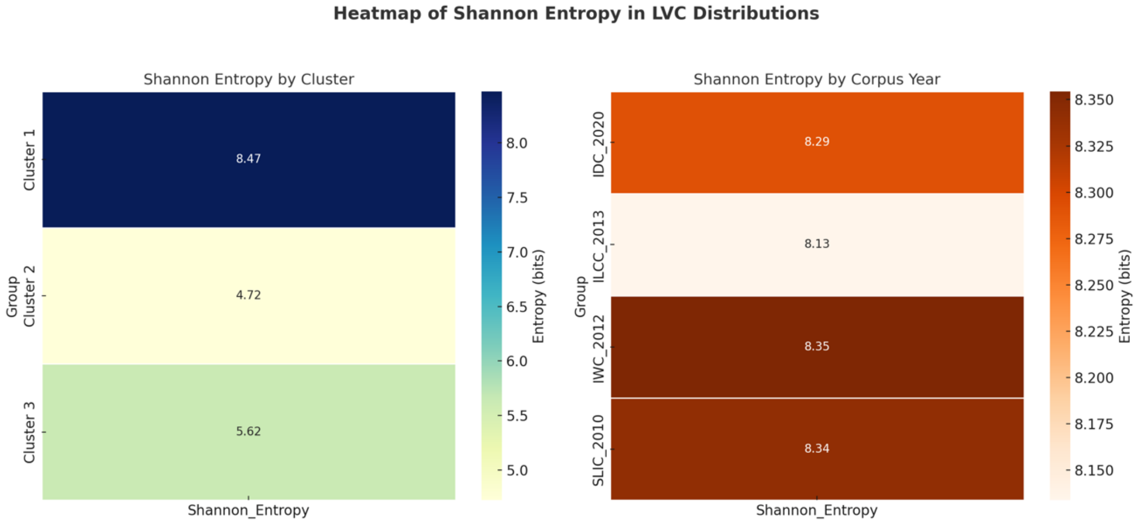


Figure 3.20: Heatmap of Shannon Entropy in LVC distribution.

When examined diachronically, entropy values across the four corpora exhibit smaller but still meaningful variation. The highest lexical unpredictability is found in IWC_2012 (H=8.35) and SLIC_2010 (H=8.34), closely followed by IDC_2020 (H=8.29). These high entropy values suggest that across these years, LVCs were used with considerable distribution balance, reflecting a lexicon characterized by frequent lexical alternation rather than fixed or formulaic usage. In contrast, ILCC_2013 exhibits the lowest entropy at 8.13, indicating a subtle but notable shift toward lexical concentration. This finding is consistent with ILCC_2013's previously observed high Yule's K (≈65.01) and zero Baayen productivity, further confirming its more repetitive and less generative linguistic structure. Although the numerical range between the highest and lowest entropy is relatively narrow (approximately 0.22), the log-scaled heatmap makes these differences visually salient, revealing how slight shifts in entropy can signal larger patterns of communicative regularity or innovation. These findings suggest that while lexical diversity remains relatively stable across time, certain corpus periods may experience micro-level contractions in expressive variability, possibly due to genre dominance or sociolinguistic standardization.

Furthermore, the integration of Shannon Entropy into this study's methodological framework enriches the understanding of LVC behavior within Indonesian. By providing a measure of lexical unpredictability, entropy offers a counterbalance to frequency-based concentration measures like Yule's K and morphological innovation indices like Baayen's P. The observed alignment between high entropy and earlier findings of lexical diversity reinforces the robustness of Cluster 1 and IWC_2012 as loci of linguistic creativity and distributional richness. Conversely, low entropy scores in Cluster 2 and ILCC_2013 corroborate their profiles as linguistically conservative segments. More broadly, entropy underscores the importance of *lexical balance* as a distinct axis of variation in the study of multiword constructions. Its sensitivity to the *relative frequency distribution,* rather than absolute type or token counts, makes it uniquely suited for capturing subtle shifts in lexical behavior across structural or temporal dimensions.

Furthermore, Simpson's Index (D) and its complement, the Gini–Simpson Index (1-D)[29], provide a statistical lens for evaluating lexical dominance and diversity in LVC distributions. While metrics such as Shannon Entropy and Yule's K capture unpredictability and concentration respectively, Simpson' D emphasizes the *probability that two randomly selected*

[29] *See* details in Greenberg (1956), Grin & Fürst (2022), Jost (2006), Keylock (2005), Simpson (1949), and Yamano (2001).

tokens will belong to the same type—an intuitive reflection of lexical skew. A high D indicates lexical dominance, where a few types account for most occurrences, whereas a high Gini–Simpson Index (approaching 1) signals a more even and diverse lexical field. In the context of LVC usage in Indonesian, these indices enable precise measurement of *distributional equity,* both across structural clusters and diachronic corpora.

The Gini–Simpson Index (0.9468) and the highest Simpson's D (0.0532), suggesting a lexicon *dominated by a few highly frequent LVCs.* Cluster 3 occupies an intermediate position with a Gini–Simpson value of 0.9766, reinforcing earlier interpretations of it as a hybrid cluster—les formulaic than Cluster 2, but not as diverse as Cluster 1. These distinctions are vividly portrayed in the Figure 3.21), where the height and labeling of each bar underscore the relative equity or concentration in lexical usage. This multidimensional alignment across metrics validates the structural segmentation of the dataset and supports the theoretical framing of LVC behavior as functionally differentiated.

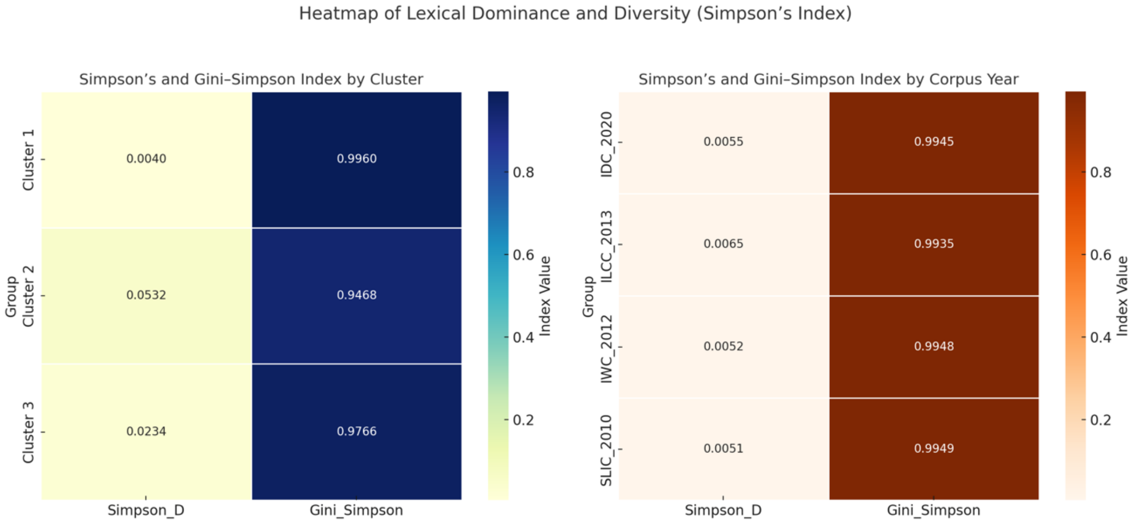


Figure 3.21: Heatmap of lexical dominance and diversity (Simpson's Index)

The diachronic analysis of lexical diversity, as captured by the Gini–Simpson Index, suggests a relatively stable equilibrium across corpus year, with minor fluctuations that nonetheless align with prior findings. The values range narrowly from 0.9949 (SLIC_2010) to 0.9935 (ILCC_2013), indicating that the probability of randomly selecting different LVC types remains consistently high across time. However, ILCC_2013 again emerges as the outlier, with the lowest Gini–Simpson score (0.9935) and the highest Simpson's D (0.0065) among the corpora. This subtle decrease in diversity mirrors earlier patterns observed in the same corpus—higher Yule's K (65.01), lower entropy (H=8.13), and zero Baayen productivity—underscoring its position as the most formulaic temporal segment. In contrast, SLIC_2010 and IWC_2012 exhibit the highest diversity, reinforcing interpretations of early LVC innovation and lexical experimentation. The plot illustrates these variations effectively, allowing even small differences to be visually grasped and statistically contextualized. These findings affirm that while overall LVC diversity in Indonesian has remained high, certain periods, particularly 2013, reflect shifts toward stylistic rigidity or institutional standardization.

3.3 Distributional analysis of Indonesian LVCs across predefined linguistic conditions

Based on the Hypothesis 3, this analysis will investigate whether naturally occurring distinctive clusters of LVCs exhibit statistically significant divergence in their distributional patterns across six predefined analytical dimensions. These dimensions are specifically designed to capture variations at the morphological, semantic, and syntactic levels (H3). Grounded in theoretical considerations concerning inherent linguistic features of LVCs and informed by empirical findings derived from clustering analyses of LVC frequency characteristics across all examined corpora, this hypothesis aims to elucidate the degree of (dis)connection between the identified clusters and their subsequent implications for a more nuanced understanding of Indonesian LVCs. Particularly, the six predefined analytical dimensions include nominal morphological features (§3.3.1), conceptual skeleton of SUBSTANCE (§3.3.2), scale of synonymity (§3.3.3), prototypicality (§3.3.4), transitivity parameter (§3.3.5), and valency frame (§3.2.6). By utilizing these six linguistic conditions, this examination provides an alternative to gain a comprehensive understanding of the distribution and usage patterns of Indonesian LVCs.

3.3.1 Distribution of LVCs in relation to nominal morphological features

This segment of analysis delves into the morphological intricacies of Indonesian LVCs, focusing specifically on their distribution across basic and affixed forms of the semantic-head noun. Basic forms are characterized by their unadorned structure, devoid of any affixation in the semantic-head noun of LVCs. Conversely, affixed forms necessitate the addition of affixes to assume their requisite form. By examining the morphological underpinnings of LVCs, this segment is anticipated to illuminate the morphological constraints and potentialities inherent in Indonesian LVCs.

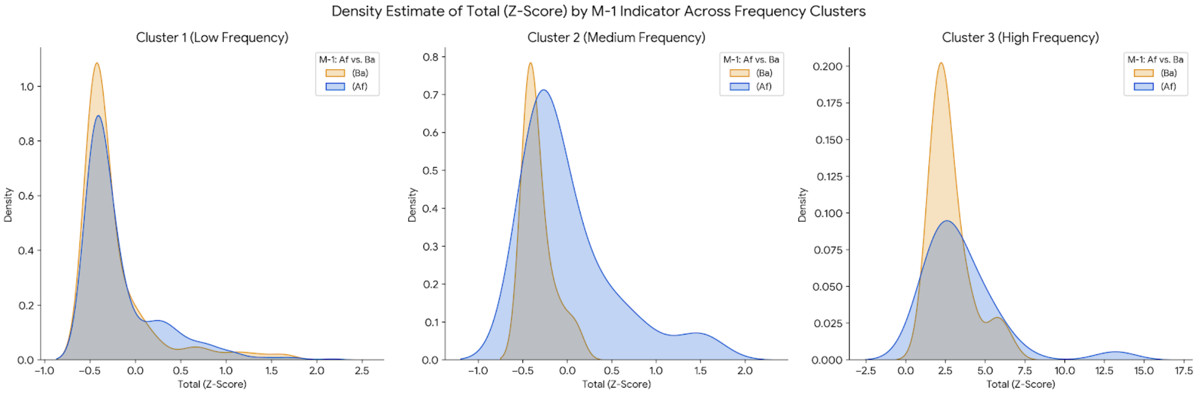


Figure 3.22: Distribution of LVCs across nominal morphological features.

Regarding the distribution of LVCs across nominal morphological features, Figure 3.22 offers a kernel density estimate (KDE) plot.[30] The presented graphical representation illustrates the KDE of total scores' LVCs (z-score) across three distinct frequency clusters, categorized by the morphological attributes of Affixed (Af) and Base (Ba) forms. The visualization approximates the probability density function of total scores' LVCs, thereby providing a smoothed depiction of distribution within each cluster. The x-axis represents the total score, while the y-axis denotes the density, reflecting the relative likelihood of observing specific scores values. The differentiation between 'Af' and 'Ba' categories is visually conveyed through shaded regions within each cluster's subplot.

Within Cluster 1, designated as the low-frequency cluster, the KDE reveals distinct distributional patterns for 'Af' and 'Ba' forms. The 'Af' category exhibits a distribution characterized by a pronounced peak around a total (z-score) of approximately -0.5, indicating a higher density of observations towards the lower end of the total score range, with a tail extending towards higher scores reaching up to 2.5. This suggests that in-low frequency

[30] Within the purview of distributional analysis of Indonesian LVCs across predefined linguistic conditions, Kernel Density Estimate (KDE) plots serve as a salient non-parametric method, statistical inference procedures applicable to data exhibiting non-normal distributional properties, for visualizing the underlying probability density function of continuous variables, such as total scores or frequency metrics (Gramacki, 2018; Silverman, 2018; Wasserman, 2004, 2006; Węglarczyk, 2018). Unlike histogram, KDE plots proffer a smoothed and continuous estimation of the data distribution, thereby mitigating the arbitrariness associated with bin selection and revealing nuanced pattern such skewness or multimodality with greater fidelity (Almodaresi *et al.*, 2017; Bowman & Azzalini, 1997; Chen, 2017; Rinaldo & Wasserman, 2010; Sheather, 2004). This facilitates a more empirical understanding of how LVC characteristics are distributed across various linguistic contexts, enabling researchers to discern subtle variations and central tendencies that might be obscured by less refined visualization techniques, enhancing the interpretability of distributional pattern within the corpus.

contexts, ‘Af’ forms are predominantly associated with lower total scores, with a minority of instances yielding higher scores. Conversely, the ‘Ba’ category demonstrates a distribution with a peak around a total (z-score) of approximately -0.2, showing a more symmetrical shape, albeit with a slight positive skew extending to around 2.0. The divergence in distributional characteristics between ‘Af’ and ‘Ba’ forms within Cluster 1 underscores the influence of morphological categorization on total score patterns at lower frequencies.

Cluster 2, representing the medium-frequency cluster, showcases a noticeable shift in the distributional patterns of total scores across the ‘Af’ and ‘Ba’ categories, although a degree of divergence persists. The ‘Af’ category exhibits a less pronounced positive skew compared to Cluster 1, with its peak density occurring around a total (z-score) of approximately -0.2, suggesting a more even distribution of total scores around the mean, yet it still maintains a distribution that leans towards lower total scores relative to the ‘Ba’ category, which peaks around a total (z-score) of approximately -0.5. The ‘Ba’ category maintains a relatively symmetrical distribution, with a slightly higher peak, indicating a greater concentration of observations around the central tendency towards higher score. While there is a reduction in the extreme skewness of the ‘Af’ category seen in Cluster 1, it is crucial to note that the two distributions remain distinct, with the ‘Af’ category clearly shifted towards higher total scores, with its distribution largely residing between -1.3 and 2.0, while the primary density of ‘Ba’ is concentrated between -1.0 and 0.5. This ongoing divergence, where the ‘Ba’ forms are associated with higher scores and the forms with lower to mid-range score, contribute to the significant association between morphological indicator and frequency cluster.

In Cluster 3, the high-frequency cluster, a degree of convergence in the distributional patterns of total scores across the ‘Af’ and ‘Ba’ categories becomes more apparent, though subtle divergences can still be observed. Both categories exhibit unimodal distributions with minimal skew, peaking at the higher end of the total score range, indicating a general tendency towards higher scores for both. However, even this cluster, a slight distributional difference remains: the ‘Ba’ category’s distribution shows a marginally higher mode and a slightly longer tail extending towards the highest scores, suggesting that while both categories are associated with high scores, the highest scores are still somewhat more characteristic of ‘Ba’ forms. While both distributions appear relatively symmetrical, the ‘Af’ category might exhibit a very slight positive skew, with a longer tail extending towards the higher scores compared to what was initially describe as ‘minimal skew’ for both.

Moreover, to substantiate these observed distributional differences, a Chi-Square test of independence was performed. The resultant significant Chi-Square statistic ($\chi^2 = 10.2946$, df = 2, p = .0058) compellingly demonstrates a statistically significant association between M-1 Indicator and Frequency Cluster. This outcome necessitates the rejection of the null hypothesis, thereby providing robust statistical evidence that the distribution of ‘Af’ and ‘Ba’ form LVCs is not uniform across frequency clusters. Consequently, the preliminary visual observations are corroborated by quantitative statistical analysis, affirming the existence of distinct distributional divergences.

3.3.2 Distribution across conceptual skeleton of SUBSTANCE

This section delves into the distributional patterns of Indonesian LVCs as they intersect with the semantic-heads of their constituent nouns. Specifically, the analysis focuses on the conceptual skeleton of SUBSTANCE of these nouns, distinguishing between concrete (Co) and abstract (Ab) entities. Concrete nouns refer to tangible objects or entities that can be perceived through the senses, such as ‘*buku*’ (lit. book) or ‘*meja*’ (lit. table). Conversely, abstract nouns represent intangible concepts or ideas, such as ‘*kebahagiaan*’ (lit. happiness) or ‘*kebebasan*’ (lit. freedom). This analysis is anticipated to shed light on the interplay between lexical semantics and grammatical constructions, uncovering the underlying mechanisms that govern the formation and usage of LVCs in Indonesian.

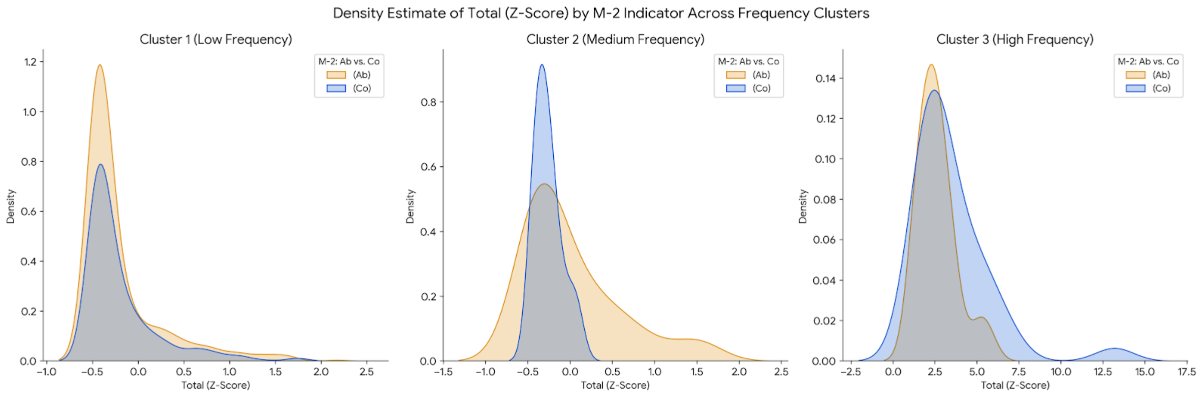


Figure 3.23: Distribution of LVCs across conceptual skeleton of SUBSTANCE.

As illustrated in Figure 3.23, the distributional characteristic of the LVCs’ total count variable, stratified by the M-2 indicator and across frequency clusters, are visually represented in a series of kernel density estimates. The clusters, designated as Cluster 1 (Low Frequency), Cluster 2 (Medium Frequency), and Cluster 3 (High Frequency), serve to categorize the data based on inherent frequency variations. The M-2 indicator delineates two distinct categories: Abstract (Ab) and Concrete (Co). Specifically, across the low-frequency cluster (Cluster 1), the density plot reveals a distribution with the highest density towards lower total (z-score) values for both 'Ab' (orange) and 'Co' (blue). The 'Ab' category peaks around a total score of approximately -0.5, exhibiting a right skew with a tail extending towards higher values up to around 2.5. Similarly, the 'Co' category shows its highest density around a total score of approximately -0.5, also displaying a right skew with a tail extending to about 2.0. This suggests that, within the low-frequency cluster, both categories are predominantly characterized by smaller total count values, although a tail extending towards higher values indicates the presence of some larger counts. A subtle difference is observable in the slightly longer tail of the 'Ab' distribution, implying a greater range of total count values for this category in this cluster.

In the medium-frequency cluster (Cluster 2), both 'Ab' and 'Co' exhibit distributions that are not similarly right-skewed distribution. The 'Ab' category shows a unimodal distribution with a peak around a total (z-score) of approximately -0.2, displaying a slight positive skew extending around 2.5. The 'Co' category, however, presents a unimodal distribution with a peak around a total (z-score) of approximately -0.4. While the density peaks are not sharper compared to the broader distributions in Cluster 1, they located at slightly different points on the lower end of the total (z-score) spectrum. This suggests that, in the medium-frequency range, 'Ab' tends to have slightly higher total scores compared to 'Co', although both still exhibits a spread towards higher values. Finally, the high-frequency cluster (Cluster 3) displays a distinct shift towards distributions with peaks at higher total (z-score) values. The 'Ab' category peaks around 2.5, while 'Co' peaks around 3.0. Both distributions in Cluster 3 show a broader spread compared to Cluster 2, indicating greater variability in total (z-score) values. They show unimodal distributions with increased variance. Notably, the 'Co' category extends to a higher maximum total (z-score) of approximately 17, indicating the presence of some exceptionally high total (z-score) values compared to 'Co', which extends to around 7.5.

Collectively, the density plots illustrate a clear modulation of the total count distribution as a function of frequency cluster. The observed shifts in skewness, kurtosis, and range underscore the heterogenous nature of the data and suggest that frequency plays a significant role in shaping the distribution of the total count variable. To substantiate the observed distributional differences in the total count variable across frequency clusters and M-2 indicator categories, a Chi-Square test of independence was conducted. The analysis yielded a statistically significant association between the M-2 Indicator and Frequency Cluster variables ($\chi^2 = 8.291$, df = 2, p = .016). This result compels the rejection of the null hypothesis, thereby providing statistical evidence that the distribution of 'Ab' and 'Co' categories is not uniform across the frequency clusters. In other words, the proportions of 'Ab' and 'Co' vary significantly depending on whether the data falls into the low, medium, or high-frequency cluster. This quantitative finding supports the preliminary visual observations derived from the kernel density estimates, affirming the existence of distinct distributional divergences as a function of both the M-2 indicator and frequency.

3.3.3 Distribution across scale of synonymity

The subsequent contextual dimension of this analysis centers on the semantic consideration, which delves into the distributional patterns of Indonesian LVCs in the scale of synonymity. Within this scale, some LVCs have a synonymous pair. Synonymous pairs, in this context, refer to LVCs that exhibit equivalent or nearly equivalent meanings while employing distinct lexical elements, i.e. a verb phrase or noun phrase. It also refers to instances where an LVC can be substituted with a synonymous morphological element without significantly altering the meaning of construction when functioned as constituent of larger construction, e.g., clause or sentence. To some extent, this semantic interchangeability highlights the inherent flexibility and versatility of LVCs in Indonesian. By examining the presence of counterpart (Cp) or absence of counterpart (Ncp) as the synonymous pairs for LVCs, this part of analysis uncovers the semantic constraints and possibilities that govern their formation and usage.

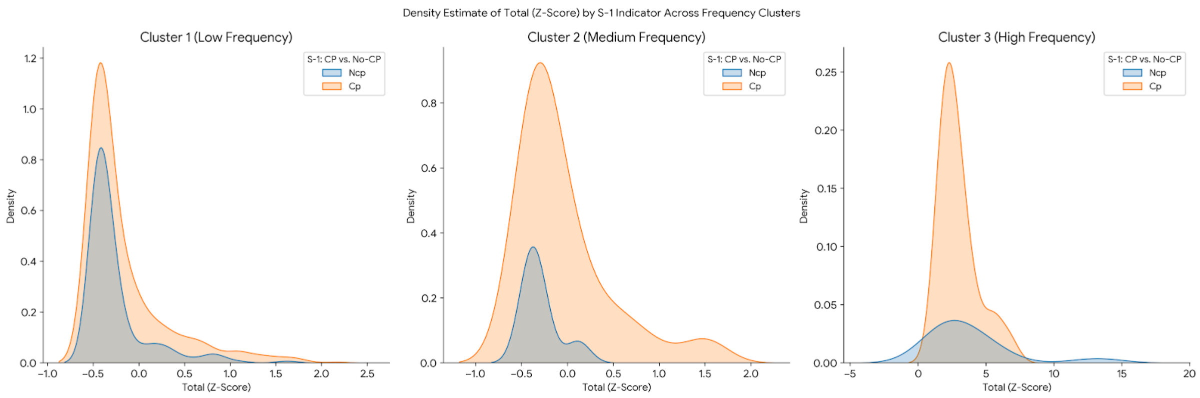


Figure 3.24: Distribution of LVCs across scale of synonymity.

As illustrated in Figure 3.24, the distribution of LVCs' total count values across frequency clusters, segmented by the S-1 Indicator, is elucidated through kernel density estimation. Within the low-frequency cluster (Cluster 1), where the mean total count is 12,325 (SD = 18,990), the density plots reveal a pronounced positive skew for both 'Cp' and 'Ncp', indicating a preponderance of lower total count values. The distribution for 'Ncp', with an overall mean of 16,826 (SD = 47,003), appears slightly more concentrated towards lower end of the scale, suggesting that instances of 'Ncp' are predominantly associated with smaller total counts in this frequency range. Conversely, the 'Cp' distribution (overall mean = 23,300, SD = 42, 257) exhibits a somewhat broader spread, implying a greater degree of variability and the occurrence of larger total count values, albeit less frequent.

In medium-frequency cluster (Cluster 2), characterized by a mean total count of 17,338 (SD = 21,096), a similar rightward skew is observed for both categories, but with a more compressed range of total count values. The density peaks are sharper and more closely aligned to the lower end of the scale, signifying a reduced dispersion and a general tendency towards lower counts for both 'Cp' and 'Ncp' in this intermediate frequency band. In addition, the high-frequency cluster (Cluster 3), showing a substantially higher mean total count of 159,032 (SD = 86.171), demonstrates a noticeable shift in distributional characteristics. While the rightward skew persists, the density curves display a more gradual decline, indicating increased variability in total count values. Notably, the 'Cp' category shows a more extended tail, suggesting the presence of some exceptionally high total count values within the high-frequency context.

Collectively, the kernel density estimates, coupled with the descriptive statistics, reveal a systematic modulation of the LVCs' total count distribution as a function of frequency cluster and S-1 indicator category. The observed variations in skewness, spread, and modal tendencies, along with the differences in mean and standard deviation across clusters, highlight the complex interplay between frequency, the S-1 indicator, and the magnitude of total counts. To substantiate these observed distributional differences, a Chi-Square test of independence was performed. The resultant significant Chi-Square statistic (χ^2 = 10.251, df = 2, p = .006) compellingly demonstrates a statistically significant association between the S-1 Indicator and Frequency Cluster. This outcome necessitates the rejection of the null hypothesis, thereby providing statistical evidence that the distribution of Cp and Ncp categories is not uniform across frequency clusters.

3.3.4 Distribution across prototypicality

The subsequent contextual dimension of this analysis centers on the semantic parameter, delving into the distributional patterns of Indonesian LVCs in relation to the prototypicality of the LVC's noun. This analysis seeks to uncover the interplay between LVCs and the semantic inherently semantic-properties of their noun element. The prototypicality of the LVC's noun refers to the distributional classification of nouns based on their prototypical and typological meanings. In this context, nouns are categorized into seven semantic classes: People (Pe), Plants (Pl), Animals (An), Material (Ma), Objects (Ob), Qualities (Qu), Action (Ac), and Processes (Pr). By examining the distribution of LVCs across these prototypicality categories, this part of analysis aims to identify the semantic constraints and possibilities that govern their usage.

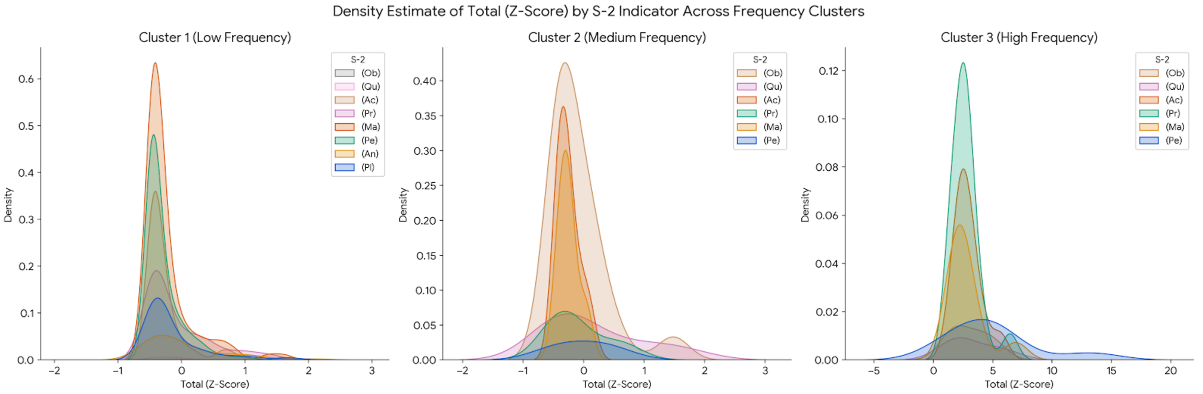


Figure 3.25: Distribution of LVCs across prototypicality.

The distribution of LVCs' total count values across frequency clusters, segmented by the S-2 Indicator, is elucidated through kernel density estimation as illustrated in Figure 3.25. Several key observations of cluster-specific patterns can be revealed. In the low-frequency (Cluster 1), the distributions for most S-2 categories exhibit a pronounced positive skew, with a concentration of data points at the lower end of the total count spectrum. The average total count values within this cluster range from 8,817 for 'People' (Pe) to 23,375 for 'Qualities' (Qu), with standard deviations indicating substantial variability (e.g., 'People' SD = 13,774; 'Qualities' SD = 27,477). This suggests that, within this cluster, instances of these categories predominantly involve smaller total count values, although the relatively large standard deviations indicate a wide spread. Notable exceptions include the 'Qualities' (Qu) and 'Objects' (Ob) categories, which display a somewhat broader distribution, suggesting a greater range of total count values, although still skewed towards lower counts ('Object' mean = 14,638, SD =

21,278). Overall, the low-frequency cluster is characterized by a high degree of variability in the distributions, with considerable overlap among the categories.

The medium-frequency cluster (Cluster 2) demonstrates a more compressed range of total count values compared to Cluster 1. The average LVCs' total-count values are generally lower, ranging from 9,590 for 'Processes' (Pr) to 28,141 for 'Material' (Ma), and the standard deviations are also smaller, indicating reduced dispersion (e.g., 'Processes' SD = 8,021; 'Material', SD = 34,649). The density curves are generally narrower, with sharper peaks, indicating a greater concentration of values within a smaller range. This suggests that, in the medium-frequency range, the total counts for most S-2 categories tend to be less variable and lower in magnitude. The 'Processes' (Pr) and 'Actions' (Ac) categories show slightly more prominent tails ('Actions' mean = 17,010, SD = 20,901), indicating the presence of higher count values, although these are relatively infrequent. Additionally, the high-frequency cluster (Cluster 3) reveals a distinct shift in distributional characteristics. The average total count values are substantially higher, ranging from 139,933 for 'Actions' to 238,381 for 'Objects', with large standard deviations reflecting increased variability (e.g., 'Actions' SD = 52,417; 'Objects' SD = 164,559). While the rightward skew persists, the density curves display a more gradual decline, indicating the occurrence of some exceptionally high total count values. The 'Objects' and 'Actions' categories display the most pronounced shift towards higher values.

Overall, the kernel density estimates highlight a systematic modulation of the total count distribution as a function of frequency cluster and S-2 indicator category. The observed variations in skewness, spread, and modal tendencies, along with the differences in mean and standard deviation across clusters and categories, underscore the complex interplay between frequency, the semantic categories represented by the S-2 indicator, and the magnitude of total counts. To substantiate the distributional differences across frequency clusters, an ANOVA was conducted. The results revealed a statistically significant effect of Frequency Cluster on LVCs' Total Count (F = 721.05, p < .001). Furthermore, a Kruskal-Wallis test revealed a significant difference in the distribution of total count values across the S-2 categories (H = 35.60, p < .001), suggesting that the semantic nature of the indicator also plays role in shaping the observed distributions.

3.3.5 Distribution across transitivity parameter

The distribution of Indonesian LVCs can be quantitatively assessed at the fundamental level of transitivity, providing a methodological framework for tracking the presence and behavior of verbal elements within these constructions. The analysis reveals a dichotomy in the distributional tendencies of these verbal elements aligning with their inherent transitivity properties. Specifically, the analysis observes two primary patterns: the utilization of inherently low-transitive verbs, which intrinsically lack the capacity to take a direct object, and the employment of inherently high-transitive verbs, which necessitate a direct object to fulfill their syntactic and semantic requirements. This distinction lays the groundwork for a more nuanced exploration of the complex interplay between lexical semantics, syntax, and argument structure within the realm of Indonesian LVCs.

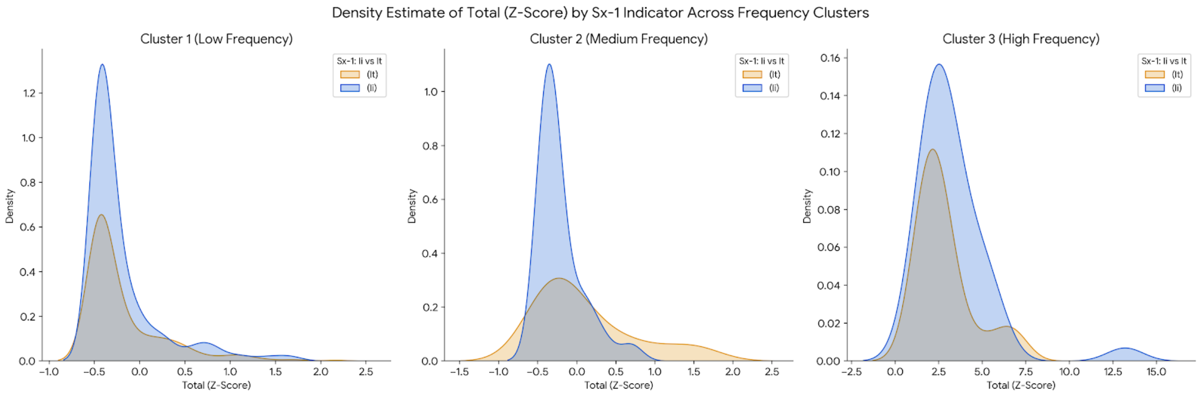


Figure 3.26: Distribution of LVCs across transitivity parameter.

As illustrated in Figure 3.26, the kernel density estimates illustrate the distribution of LVCs' total-count values across three frequency clusters, categorized by the Sx-1 indicator, which differences between *inherently intransitive* ('Ii') and *inherently transitive* ('It') verb constructions. Several cluster-specific observations can be revealed. Within the low-frequency cluster (Cluster 1), both 'Ii' and 'It' categories exhibit a right-skewed distribution, with a concentration of data points at the lower end of the total count spectrum. The mean total count for 'Ii' constructions is 12,811 (SD = 19534), while for 'It' constructions, it is 12,046 (SD = 18,683). The similarity in means suggests that, in the low-frequency cluster, there is no substantial difference in the central tendency of total counts between the two constructions types. However, the relatively large standard deviations for both categories imply considerable variability in total count values. The density curves show a sharp peak at the lower end, tapering off gradually, which is characteristic of a positive skew.

In the medium-frequency cluster (Cluster 2), a distributional shift is observed. The mean total count for ‘Ii’ constructions increases to 25,769 (SD = 27,649), and for ‘It’ constructions, it decreases to 11,405 (SD = 12,350). This divergence in means indicates a potential influence of verb transitivity on total count in the medium-frequency range, with ‘Ii’ constructions exhibiting higher average counts and greater variability. The ‘It’ distribution appears more concentrated towards the lower end, while the ‘Ii’ distribution shows a broader spread. Additionally, the high-frequency cluster (Cluster 3) displays the most pronounced differences. The mean total count for both categories is substantially higher compared to the other clusters: 145,810 for ‘Ii’ (SD = 70,662) and 166,589 for ‘It’ (SD = 94,030). The ‘It’ constructions demonstrate a slightly higher average total count and greater variability in this cluster. The density curves in Cluster 3 are less skewed, indicating a more even distribution of total count values, although a peak remains discernible at the lower end.

Overall, the kernel density estimates reveal a systematic variation in the distribution of total count values across frequency clusters, modulated by the Sx-1 indicator. The observed variations in skewness, spread, and modal tendencies, along with the differences in mean and standard deviation across clusters, underscore the interplay between frequency clusters, the syntactic categories represented by the Sx-1 indicator, and the magnitude of total counts. To further substantiate these observations, a Chi-Square test of independence was performed to assess the relationship between frequency cluster and Sx-1 indicator. The analysis yielded a non-significant result ($\chi^2 = 0.436$, df = 2, p = .804), indicating that the distribution of ‘Ii’ and ‘It’ categories is not significantly different across the frequency clusters. This suggest that while the *magnitude* and *distributional shape* of total counts vary across clusters, the *proportions* of inherently intransitive and inherently transitive constructions are relatively consistent across these frequency bands.

3.3.6 Distribution across valency frame

The subsequent syntactic parameter under scrutiny in the analysis of Indonesian LVCs pertains to the domain of valency. In contrast to bare verbs, which characteristically exhibit a rigid and well-defined valency spectrum, LVCs manifest a divergent tendency. This divergence stems from their inherent nature as compositional units, wherein the amalgamation of a light verb and a noun engenders a novel semantic entity. Consequently, the valency of an LVC cannot be sole predicted from the valency of its constituent verb. To capture this complexity, the current investigation examines the distributional patterns of LVCs within two distinct valency paradigms: those exhibiting a fixed valency, mirroring the behavior of their constituent light verbs, and those demonstrating a more flexible valency, influenced by the semantic and syntactic properties of the incorporated noun. This binary classification provides a framework for discerning the intricate interplay between lexical semantics, syntactic structure, and argument realization within the realm of Indonesian LVCs.

As illustrated in Figure 3.27, the kernel density estimates visualize the distribution of total count values across three frequency clusters, categorized by the Sx-2 indicator, which distinguished between fixed valency (‘Fi’) and flexible valency (‘Fe’) LVCs. Several cluster-specific observations can be revealed. Within the low-frequency cluster (Cluster 1), both ‘Fi’ and ‘Fe’ constructions exhibit a right-skewed distribution, indicative of a preponderance of lower total count values. The mean total count for ‘Fe’ constructions is 13,997 (SD = 20,921), while for ‘Fi’ constructions, it is 11,577 (SD = 18,029). These statistics suggest a slightly higher average total count for ‘Fe’ constructions, though the substantial overlap in distribution and comparable standard deviations indicate considerable variability within both categories. The density curves are characterized by a sharp ascent from the origin, followed by a gradual decline, typical of a positive skew.

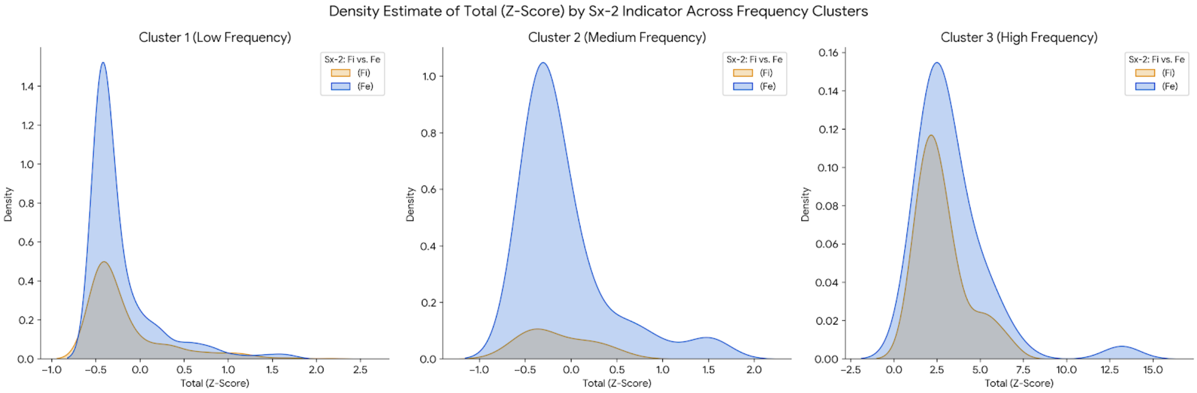


Figure 3.27: Distribution of LVCs across valency frame.

In the medium-frequency cluster (Cluster 2), the mean total for ‘Fi’ constructions (17,710, SD = 21,782) is again higher than that for ‘Fe’ constructions (14,287, SD = 15,787). However, the overall range of total count values is compressed relative to Cluster 1, and the standard deviations are somewhat reduced, suggesting less variability. The density curves display a more peaked shaped, indicating a concentration of data points within a narrower range of total count values. In addition, the high-frequency cluster (Cluster 3) demonstrates a marked increase in the magnitude of total count values for both categories. The mean total count for ‘Fi’ constructions reaches 167,907 (SD = 97,027) and for ‘Fe’ constructions, it is 143,503 (SD = 62,118). The ‘Fi’ constructions continue to exhibit a higher average total count and greater variability. The density curves in this cluster show a less pronounced skew, indicating a more even distribution of total count values, although a discernible peak remains at the lower end.

Overall, the kernel density estimates reveal a systematic influence of frequency cluster on the distribution of LVCs’ total count values across fixed and flexible valency constructions. While ‘Fi’ constructions consistently exhibit higher average total counts, the distributional shapes and the degree of variability change considerably across the frequency spectrum. The high-frequency cluster is distinguished by substantially larger total count values and a broader spread of the data. To further validate these observations, a Chi-Square test of independence was performed to assess the relationship between frequency cluster and Sx-2 indicator. The test yielded a statistically significant result (χ^2 = 9.339, df = 2, p = .009), indicating that the distribution of the fixed and flexible valency constructions is not uniform across the frequency clusters. This suggests that the frequency of occurrence is associated with variations in the proportion of these constructions’ types.

3.4 Discussion

3.4.1 A short discussion of the three central hypotheses

The central hypotheses (*see* §3.0) posited in this chapter find substantial support in the empirical evidence gleaned from the corpus-based analysis of Indonesian LVCs. First, the examination into the cross-corpus frequency of linguistically defined LVCs, as detailed in the preceding section, yields several key insights relevant to Hypothesis 1. This hypothesis posited that, contrary to initial expectations of significant divergence due to inherent compositional disparities between the hypothetical and genuine LVC datasets, specific linguistically well-defined LVC types would exhibit statistically significant correlations in frequency across the four corpora under investigation, when controlled for genre and time period. The results of the correlation analyses provide support for this hypothesis. The observed correlations in the frequency of certain LVC type across the ILCC, SLIC, IDC, and IWC corpora suggest that, despite the differences in corpus composition and the temporal variations, there are underlying patterns of LVC usage that transcend corpus-specific idiosyncrasies. Specifically, the analysis of 'true light verb constructions', 'vague action verb constructions', 'eventive light verb constructions', and 'stative light verb constructions' reveals that some of these categories demonstrate a degree of consistency in their occurrence patterns.

However, it is crucial to acknowledge that the correlations are not uniform across all LVC types. The observed variations in the strength and significance of the correlations across the different categories indicate that the factors influencing LVC usage are complex and multifaceted. While genre has been controlled for, the influence of time period, as acknowledged in the analysis, requires careful consideration. The temporal distance between the corpora (ILCC in 2013, SLIC in 2010, IDC in 2020, and IWC in 2012) may contribute to some of the observed variations. For instance, any significant differences between SLIC (2010) and IDC (2020) might be indicative of diachronic change in LVC usage. Furthermore, the "inherent compositional disparities" between the hypothetical and genuine LVC datasets, as initially anticipated, play a role in shaping the observed frequency patterns. The process of identifying and categorizing LVCs, even with linguistically well-defined criteria, involves a degree of interpretation that may introduce variability. The hypothetical dataset, derived from theoretical considerations, may differ systematically from the genuine dataset, which reflect introspective original-perspective from native Indonesian. This difference could explain why not all LVC types exhibit strong correlations across the corpora.

Second, the investigation into the frequency distribution of LVCs as multiword expressions addresses Hypothesis 2, which predicts a deviation from a strict Zipfian distribution, particularly in the lower frequency range. The analysis, employing a combination of visual inspection (log-log plot), regression analysis, and goodness-of-fit tests, provides evidence supporting this hypothesis. While Zipf's Law, which posits an inverse relationship between word's frequency and its rank, is observed in many linguistic contexts, the behavior of LVCs presents a notable departure. The log-log plot analysis, a primary tool for visualizing Zipfian distributions, reveals that LVC frequency distribution does not conform to the expected linear pattern with a slope of -1. Instead, the plot demonstrates a curve, indicating a higher-than-predicted frequency of lower-ranking LVCs.

This deviation is further explained is explained by the inherent characteristics of LVCs. Unlike single-word lexical items, LVCs are multiword expressions with intricate syntactic structures and nuanced semantic functions. Their meaning is not simply the sum of their individual components but rather a composite of the light verb and the following constituent. This complexity restricts the variability and predictability of their occurrence. The complex interplay of syntactic and semantic factors governing LVC usage contributes to their non-Zipfian distribution. The light verb, while frequent in isolation, constraints the selection of its complement, resulting in a limited range of possible LVC combinations. Furthermore, the semantic opacity of some LVCs, where the meaning is idiomatic or highly context-dependent, further reduces their likelihood of appearing with high frequency. In deduction, the findings indicate that LVCs, as multiword expressions, exhibit a frequency distribution that deviates from the pattern observed for single-word lexical items, as described by Zip's Law. The complex syntactic and semantic properties of LVCs play a crucial role in shaping their distribution. This highlights the importance of considering the multiword nature of linguistic units in frequency-based analyses.

Third, the distributional analysis of LVCs across six predefined analytical dimensions directly addresses Hypothesis 3. This hypothesis posits that naturally occurring distinctive cluster of Indonesian LVCs exhibit statistically significant divergence in their distributional patterns across these dimensions, which are designed to capture variations at the morphological, semantic, and syntactic levels. The findings of this analysis, particularly when considered in the light of the three LVC clusters (Q1: Low Frequency, Q2: Medium Frequency, and Q3: High Frequency) derived from K-means clustering, and the KDE plots illustrating their distribution across the analytical dimensions, provide a nuanced perspective of LVC behavior. This stratification facilitates a more nuanced exploration of the factors influencing LVC distribution, enabling comparisons and contrasts across different levels of frequency, underscoring the crucial role these factors play in meaning construction (cf. Caro & Arús-Hita, 2020; Mattissen, 2023; Pompei, 2023). The K-means clustering effectively grouped LVCs based on their frequency characteristics, revealing inherent differences in how these constructions are employed within the language. This clustering, combined with the examination of the six analytical dimensions and the KDE plots, allows for a more granular understanding of LVC distribution. The KDE plots provide a visual representation of the probability density of LVCs across each dimension, highlighting both central tendencies and the spread of the data for each cluster. For instance, the analysis reveals that LVCs within Cluster Q1 (Low Frequency) exhibit distinct distributional patterns compared to those in Cluster Q3 (High Frequency), and this is further clarified by the KDE plots.

Specifically, LVCs in Q1, characterized by their infrequent occurrence, tend to show a preference for specific morphological feature. The KDE plots for this cluster indicate a distribution skewed towards more complex morphological structures, such as affixed forms of the semantic-head noun. This suggests that these less frequent LVCs might serve more specialized grammatical or semantic functions, requiring more explicit morphological marking. This contrast with LVCs in Q3, which, due to their higher frequency, often occur in more basic, less morphology complex forms. The KDE plots for Q3 show a distribution concentrated around simpler morphological structures, suggesting that frequently used LVCs may undergo a process of simplification or standardization over time, al also evidence by the KDE. This finding elucidates the distinct morphological and morpho-semantic characteristics of the light verb, corroborating previous observations regarding the inherent flexibility of LVCs (e.g. Bouveret, 2021; Jezhek, 2023). This semantic-head noun fulfills two primary functions (cf. Mlac & Tournadre, 2021; Toluspayeva *et al.*, 2024). Firstly, it contributes significantly to the overall semantic content of the construction, enriching the bleached semantics of the light verb itself. The inclusion of this noun serves to clarify and specify the intended meaning, rendering the sentence semantically complete. Secondly, the morphological and semantic attributes of the semantic-head noun are indispensable in determining the well-formedness of an LVC.

Furthermore, the analysis of the conceptual skeleton of SUBSTANCE reveals that the different clusters exhibits affinities for different morpho-semantic categories. LVCs in Q1, as illustrated by the KDE plots, might be associated with more abstract or specialized semantic domains, showing a narrower and more peaked distribution across the SUBSTANCE dimension. In contrast, those in Q3 are linked to more concrete and general concepts, with KDE plots indicating a broader and more dispersed distribution. Similarly, the scale of synonymity varies across the clusters, with Q1 LVCs potentially displaying a narrower range of synonyms, indicating a more precise or restricted meaning, compared to the broader synonymy range often observed in Q3 LVCs, a trend that is also supported by the KDE plots. This association, in turn, is influenced by several factors, including universal lexico-semantic principles (Goddard, 2001), syntagmatic relations (Langacker, 2022), and the overall grammatical structure (Langacker, 2020).

The dimensions of prototypicality, transitivity parameter, and valency frame also contribute to the observed divergence. Q1 LVCs, due to their lower frequency and specialized usage, might exhibit lower prototypicality scores, indicating a less central or more peripheral role in the LVC category. The KDE plots for these dimensions would show a distribution shifted away from the prototypical values. They may also show a preference for specific transitivity patterns or valency frames, reflecting their unique syntactic behavior, as visualized in the KDE plots. In contrast, Q3 LVCs, being more frequent and widely used, tend to be more prototypical and exhibit greater flexibility in terms of transitivity and valency. The KDE plots for Q3 would show a distribution concentrated around the prototypical values and a wider spread across different transitivity and valency options. Thus, this finding has shown that the distribution of Indonesian LVC is not static, but rather shows a spectrum of interpretations that depend on the semantic context and inherent syntactic properties embedded in these elements (cf. Hellan, 2023; Pompei & Piunno, 2023).

In conclusion, these findings highlight the importance of considering both frequency and linguistic properties when analyzing LVCs. The statistically significant differences observed across the six predefined dimensions, when viewed through the lens of the Q1, Q2, and Q3 clusters, and further elucidated by the KDE plots, suggest that LVCs are not a homogenous category. Instead, they represent a spectrum of constructions with varying degrees of frequency, complexity, and semantic specificity. In conclusion, the analysis confirms that naturally occurring distinctive clusters of LVCs, as identified by K-means clustering, exhibit statistically significant divergence in their distributional patterns across the six predefined analytical dimensions.

3.4.2 On the asymmetrically bound structures of Indonesian LVCs

In constructions where predictability and co-selection are central to meaning, specific law of language, i.e., conditional entropy for lexical analysis $H(Y|X)$, provides an empirical framework for analyzing the dependency structure between linguistic element.[31] Unlike Shannon entropy, which measures the overall unpredictability of a distribution, conditional entropy quantifies how much uncertainty remains about one variable—such as a noun—once the value of another variable—such as a verb—is known. In the context of LVCs, this allows for an empirical examination of the *collocational binding strength* between verb and noun components. Specifically, by calculating H(Noun | Verb) and H(Verb | Noun), we can assess whether LVCs are structurally symmetrical or asymmetrical in their collocational tendencies. This approach complements existing lexical metrics such as Yule's K, entropy, and Simpson's D by modeling

[31] PyInform is a package specifically designed for information-theoretic measures on discrete probability distributions. It provide a dedicated function `pyinform.shannon.conditional_entropy(p_xy, p_y, b=2.0)` to compute conditional entropy.

the *internal configurational rigidity or flexibility* of LVCs. it is particularly useful in languages such as Indonesian, where LVCs serve both grammatical and stylistic functions, and where certain light verbs (e.g., the verb in Cluster 3: High Frequency) are assumed to dominate semantically bleached constructions.

The results of conditional entropy analysis reveal a *marked asymmetry* in the internal dependency structure of LVCs (*see* Figure 3.28). The conditional entropy of the noun given the verb—H(Noun | Verb)—was calculated at 2.70 bits, indicating a relatively high degree of lexical variability: knowing the verb does not strongly predict which noun will follow. In contrast, the entropy of the verb given the noun—H(Verb | Noun)—was only 0.46 bits, suggesting a significant restriction in verb choice once the noun is specified. This implies that certain nouns co-occur consistently with particular verbs, forming *semantically entrenched templated* that exhibit strong selectional preference. These values suggest that, within Indonesian LVCs, nouns are the *collocational anchors,* while verbs exhibit greater contextual plasticity.[32] This pattern aligns with the grammaticalization hypothesis, where light verbs lose semantic specificity and instead serve syntactic or aspectual functions, allowing a broad range or nominal complements. The visual representation through horizontal bar plot further accentuates this asymmetry, showing that the predictability of the verb slot is substantially higher than that of the noun slot. Such a pattern points to the entrenchment of certain noun–verb pairings, and the lexical creativity of verbs in accommodating semantic content.

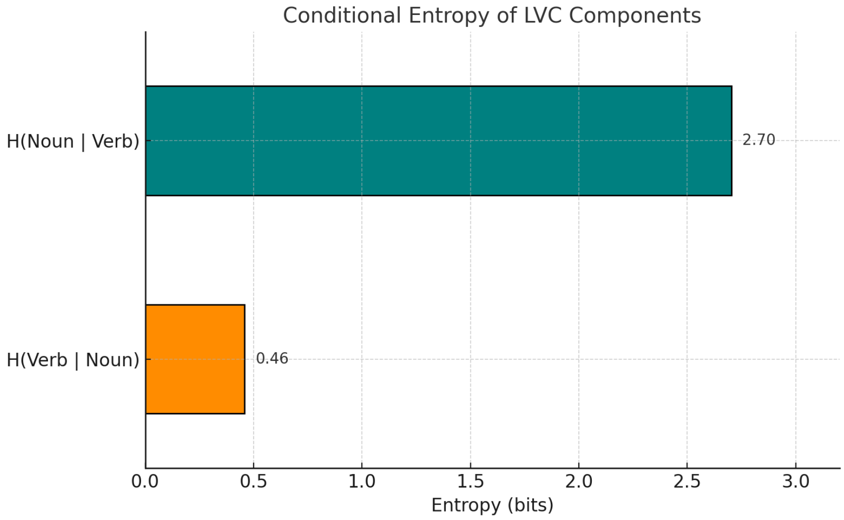


Figure 3.28: Conditional entropy of LVCs components.

[32] A comprehensive analysis of verbal elements is presented in Chapter 4, while nominal elements are discussed in detail in Chapter 5.

This entropy asymmetry provides a quantitative foundation for characterizing the internal architecture of LVCs. The lower entropy in the verb slot given a noun suggest that many nouns are semantically tethered to a specific verb, reinforcing the notion that LVCs in Indonesian are often built around noun-governed collocational templates. This is consistent with prior linguistic theories that emphasize the unidirectionality of selectional constraints, wherein nouns evoke particular light verbs to realize specific eventive meanings (e.g., *'mengambil keputusan*' (to make a decision), '*memberikan bantuan*' (to give assistance)). From a morphosyntactic perspective, the result suggests that noun components serve as *semantic heads* in these constructions, while verbs increasingly function as *light lexical scaffolding.* The high entropy associated with nouns, meanwhile, implies that verbs are relatively unconstrained and can combine with a wide array of noun complements to accommodate stylistic, discursive, or genre-specific variation. This duality between *rigidity* and *flexibility* mirrors broader dynamics observed in multiword expressions and formulaic sequences across languages, where one constituent exhibits high collocational stability while the other contributes to variation. Thus, conditional entropy serves not only as a statistical metric but as a structural diagnostic for uncovering hidden pattern in LVC composition.

The inclusion of conditional entropy in this study substantially deepens the theoretical model of LVCs by introducing a combinatorial perspective on predictability. While frequency-based and type-based measures such as Yule's K and Shannon entropy provide essential insights into lexical richness and repetition, conditional entropy addresses a more fine-grained linguistic question: how do the components of a construction interact to constrain or allow variability? The entropy asymmetry identified here support the view that Indonesian LVCs are asymmetrically bound structures, with nouns exerting greater control over lexical selection.

3.5 Resume

Cluster 3 presents analysis of the frequency and distributional behavior of Indonesian LVCs across four language corpora, advancing three central hypotheses through corpus-based and statistical methodologies. Initial analysis evaluates LVC frequency across both hypothetical and genuine datasets, revealing inter-corpus disparity and the emergence of three natural clusters (low, medium, high frequency) through K-means clustering. These groupings confirmed Hypothesis 1 and provided a foundational framework for subsequent analysis. A Spearman correlation test conducted on the ranked frequency values across corpora produced a statistically significant results (rs = 0.891, p < 0.001), indicating strong consistency in LVC frequency ordering and supporting the stability of distributional patterns over time and across genres.

Further analysis of Hypothesis 2 demonstrated that while the LVC dataset significantly deviates from a pure Zipfian distribution—especially in the long tail of rare constructions—it more closely to the Zipf–Mandelbrot law. The dataset also aligns with the Menzerath–Altmann law, with the most robust model achieved under a morpheme-based framework. Complementary insights were obtained through the application of Heap's Law, which confirmed a sublinear vocabulary growth pattern in LVC types relative to corpus size, and Baayen's productivity measure, which revealed that LVC productivity is skewed toward a narrow set of morphologically conservative clusters. Additional model using the Good-Turing estimator and Shannon entropy highlighted the limited predictability of unseen types and unevenness of structural distribution. Additional modeling of lexical concentration across time was performed using Yule's K, which indicated subtle shifts in the concentration of LVC types across corpus years. This diachronic trend was further supported by KL Divergence, which measured the degree of lexical drift between temporal segments. Notably, a dedicated analysis juxtaposed Altmann's (1967) dictionary-based lexical model of Indonesian with empirical metrics from contemporary LVC usage. The comparison revealed sharp contrasts in morpheme length, phonological density, frequency clustering, and correlation strength—thereby challenging assumptions based on lexicon-based models.

Final analysis addressed Hypothesis 3 by demonstrating that three frequency-based LVC clusters divergence significantly across six morpho-semantic-syntactic parameters. These findings confirm that Indonesian LVCs form a structurally asymmetric, distributionally layered system. High-frequency constructions tend to be morphologically simple, semantically concrete, and syntactically economical, while lower-frequency LVCs are characterized by morphological richness, semantic abstraction, and restricted syntactic mobility. Collectively, the findings in this chapter underscore the theoretical claim that LVCs in Indonesian are not structurally homogenous but exhibit gradient variation driven by usage frequency, morpho-semantic load, and contextual specialization.

Chapter 4

Verb element within Indonesian LVCs

4.0 Introduction

In the context of Indonesian LVCs, the verb element, in conjunction with a nominal constituent, engenders semantically cohesive phrasal units. This chapter undertakes a detailed analysis of the verbal component within such constructions, identifying two principal tendencies that characterize the light verb category in Indonesian, as visually represented in Figure 4.1, which illustrates the distribution between what shall be termed *true light verbs* (TLVs) and *vague action verbs* (VAVs). The first of these tendencies will be explored in §4.1. Furthermore, the second identified tendency will be the focus of §4.2.

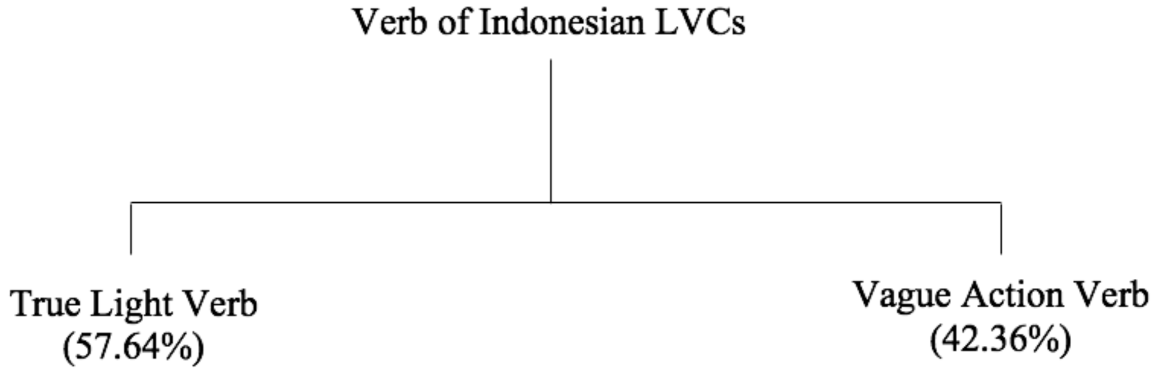


Figure 4.1: Hierarchical classification of verb elements in Indonesian LVC, with percentage division across categories.

To further explore the characteristics and distribution of these two verb categories, the analysis employs KDE plots to analyze their frequency across different levels of usage. In detail, KDE plot illustrates the distribution of total values for two verb categories, TLVs and VAVs, across three distinct frequency clusters as illustrated in Figure 4.2. These clusters—Low, Medium, and High—represent a gradient of LVCs frequency, allowing for an examination of how verb type influences distribution across varying levels of usage. The visualization reveals differences in both the central tendency and spread of the total count distributions for TLVs and VAVs constructions within and across there frequency clusters, suggesting a nuanced relationship between verb type and frequency of occurrence.

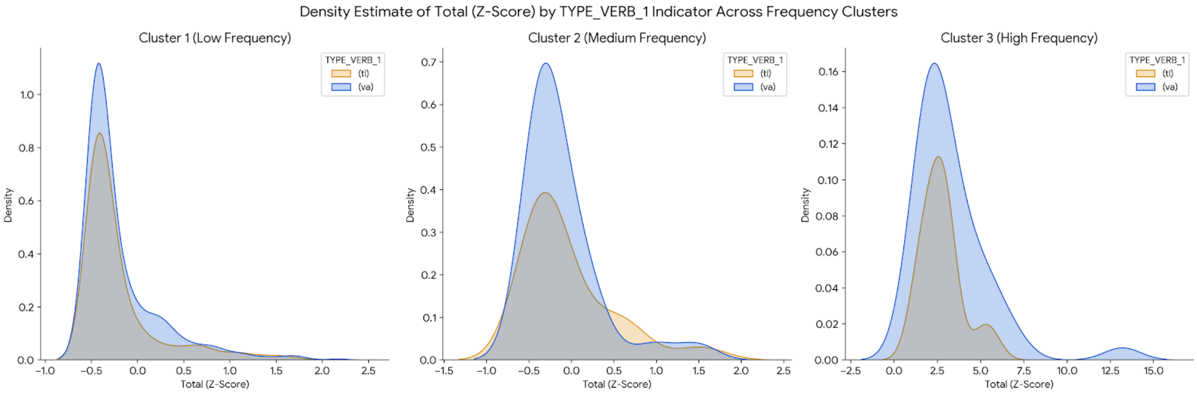


Figure 4.2: Distribution of TLVs and VAVs across frequency's clusters.

In Low Frequency cluster (Cluster 1), both TLVs and VAVs exhibit a string positive skew, with the majority of data points clustered the lower end of the total count range. This indicates that, within the realm of infrequent LVC usage, both verb types are predominantly characterized by small total count values. However, the TLVs show a slightly longer tail, suggesting that while most TLVs have low counts, there is a greater possibility of encountering some TLVs with moderately higher counts compared to VAVs. Thereafter, the Medium Frequency cluster (Cluster 2) demonstrates a marked shift in the distribution of VAVs. While still positively skewed, the VAVs distribution becomes more platykurtic, with a wider spread of total count values compared to the Low Frequency cluster. This suggests that as we move towards more frequently used verbs, VAVs exhibit greater variability in their occurrence. In contrast, TLVs maintain a more leptokurtic distribution, indicating a narrower range of total count values concentrated towards the lower end, though with a slight shift towards higher values compared to the Low Frequency cluster.

Within the High Frequency cluster (Cluster 3), both TLVs and VAVs display a further shift towards higher total count values. However, the VAVs exhibit a more pronounces platykurtic distribution, indicating that high-frequency VAVs encompass a broader range of total count values, including extremely high values. The TLVs, while also present in the high-frequency range, show a distribution that, while skewed, is concentrated at the lower end of the high-frequency spectrum. Overall, the density plot reveals a clear trend: as we more from low to high-frequency clusters, VAVs tend to exhibit a wider spread of total count values, particularly in the Medium and High Frequency clusters. This suggests that vague action verbs are used across a broader range of contexts and with greater variability in their frequency compared to true light verbs, which tend to be concentrated in the low-frequency and the lower end of the high-frequency clusters.

4.1 True light verbs

Within the landscape of Indonesian LVCs, a primary category emerges: TLVs. Drawing upon the theoretical framework outlined in Chapter 1, TLVs in Indonesian LVCs are characterized by the retention of a discernible, albeit often diminished, semantic content. As established in the preceding theoretical discussion, TLVs typically exhibit a propensity for *negative* [-] values across most of features such as durativity, frequentativeness, iterativeness, or intensiveness, indicating a limited inherent temporal specification. Their primary function lies in providing grammatical support to the noun, thereby facilitating its integration within the clausal structure.

4.1.1 Optimal TLVs

This subsection is dedicated to presenting an analysis of the optimal and suboptimal TLVs. Taking advantage of the frequency and distribution analysis conducted in Chapter 3, coupled with the implementation of a Naïve Bayes (hereafter: NB)[33], we can discern distinct tendency of their optimal type-pairing rate. The optimal TLVs demonstrate an optimal combinatorial capacity with nouns within LVC contexts, suggesting a high degree of semantic and syntactic

[33] See also other implementations in Abraham *et al.*, (2024), Jurafsky & Martin (2009), Mahowald *et al.* (2021), Raschka (2014), and Suzgun *et al.* (2022)

compatibility (*see* Table 4.1). Conversely, the suboptimal TLVs exhibit minimal pairings with nouns, indicating a more restricted range of collocational possibilities (*see* Table 4.2).

Table 4.1: Sample list of optimal TLVs.

No	**TLV**	**Optimal Type-Pairing Rate**	**Sample LVCs**	**PMW**
1.	*melakukan*	141	*melakukan perjalanan* ‘take a trip’	7864.954
2.	*memberikan*	120	*memberikan informasi* ‘give information’	11273.390
3.	*membuat*	53	*membuat jawaban* ‘make an answer’	2509.177
4.	*mengalami*	34	*mengalami kerusakan* ‘to suffer damage’	8554.266
5.	*memberi*	31	*memberi peringatan* ‘give warning’	1469.986
6.	*mengambil*	21	*mengambil keputusan* ‘take a decision’	15345.272
7.	*menjadi*	18	*menjadi saksi* ‘to bear witness’	5126.291
8.	*memiliki*	16	*memiliki pengaruh* ‘have influence’	3661.278
9.	*menyampaikan*	16	*menyampaikan pidato* ‘deliver a speech’	1766.324
10.	*mengadakan*	12	*mengadakan pertemuan* ‘have a meeting’	4454.910

Specifically, Table 4.1 presents a ranking of the ten most frequent TLVs found within Indonesian LVCs across all datasets in this study. It provides insights into the combinatorial tendencies of these verbs, highlighting their productivity and the diversity of LVCs they participate in. Several key observations as follows. Table 4.1 presents a result of the maximum observed pairings of each TLV with nouns to form LVCs. Verb ‘*melakukan*’ stand out with the highest optimal pair rate of 141, indicating their remarkable versatility in combining with various nouns to form LVCs. Table 4.1 also provides illustrative examples of LVCs formed with each TLV. These examples showcase the semantic diversity of the resulting constructions, ranging from concrete actions like ‘*menyampaikan pidato*’ (deliver a speech) to more abstract notions like ‘*memiliki pengaruh*’ (have influence).

(4.1) *Kamu tentu pernah* ***melakukan perjalanan*** *misalnya karya wisata.*
You certainly already do journey for example work tour
‘You’ve certainly gone on a journey before, for example, a school trip.’

(4.2) *Saya hanya ingin* ***memberikan informasi.***
I only want to give information
‘I just want to give (some) information.’

(4.3) *Kamu terkadang sulit* ***mengambil keputusan.***
you sometimes difficult take decision
‘You sometimes have difficulty making a decision.’

(4.4)	*Rektor*	***menyampaikan***	***pidato***	*Laporan*	*Program*	*kerja.*
	rector	convey	speech	Report	Program	Work

‘The rector delivered a speech on the Work Program Report.’

(4.5)	*Dua*	*lembaga*	*ini*	***memiliki***	***pengaruh***	*langsung*	*terhadap*	*masyarakat.*
	two	institutions	this	have	influence	direct	toward	society

‘These two institutions have a direct influence on society.’

In details, based on the sample of LVCs in (4.1) to (4.5), verbs such as ‘*melakukan*’ (to do)*,* ‘*memberikan*’ (to give)*,* ‘*mengambil*’ (to take)*,* ‘*menyampaikan*’ (to deliver)*,* dan ‘*memiliki*’ (to have) exhibit a marked tendency toward [-] values across five aktionsart dimensions—momentariness, durativity, frequentativeness, iterativeness, and intensiveness. This does not imply the absolute absence of all temporal specifications, but rather a recurrent attenuation of aspectual force in the verb element. For example, ‘*melakukan perjalanan*’ (to make a journey), the durativity is contributed primarily by the noun *perjalanan,* not by the verb *melakukan.* Similarly, in ‘*memberikan informasi*’ (to give information) or ‘*mengambil keputusan*’ (to make decision), the core semantic load is borne by the noun, with the verb functioning as a grammatical trigger. Even in ‘*menyampaikan pidato*’ (to deliver speech), which might evoke [+duration] or [-iterative] interpretations, there properties arise from the event type encoded in the noun, not from the verb itself. As such, the verbs consistently exhibit tendency toward aktionsarten neutrality, enabling flexible deployment across contexts. This pattern supports their classifications as TLVs: structurally permissive, semantically bleached, and functionally dependent on the noun temporal and conceptual specification.

Table 4.2: Extended-sample list of optimal TLVs.

No	TLV	Optimal Type-Pairing Rate	Sample LVCs	PMW
1.	*menghadapi*	7	*menghadapi tantangan* ‘be faced with a challenge’	6709.940
2.	*melayangkan*	6	*melayangkan surat* ‘to write a letter’	3319.987
3.	*membuka*	6	*membuka peluang* ‘make a chance’	8057.925
4.	*mendapat*	6	*mendapat dukungan* ‘receives support’	9336.899
5.	*menjalani*	6	*menjalani perawatan* ‘undergo treatment’	6449.113
6.	*merasakan*	6	*merasakan dampak* ‘take influence’	1023.470
7.	*membangun*	5	*membangun komunikasi* ‘to get in touch’	1004.836
8.	*mencapai*	5	*mencapai ketinggian* ‘get high’	722.813
9.	*mengajukan*	5	*mengajukan pertanyaan* ‘asks a question’	3991.770
10.	*mengeluarkan*	5	*mengeluarkan suara* ‘make a noise’	1574.859

Moreover, concerning variability in productivity, Table 4.2 highlights the variability in the productivity of different TLVs. While some verbs demonstrate a high degree of combinatorial flexibility, others appear to be more selective in their pairings with nouns. This variability reflects the semantic and syntactic constraints governing the formation of LVCs, as well as the specific roles these verbs play in expressing various actions or states. Based on Table 4.2, we can observe that the optimal pairing rate varies considerably across the ten TLVs, ranging from 5 to 7. This suggests that some TLVs, such as '*menghadapi*' (to face), exhibit greater combinatorial potential and can form meaningful LVCs with a wider range of nouns, while others, like '*membangun*' (to build) or '*mencapai*' (to reach), might have more restricted combinatorial possibilities. For instance, consider these following examples of less frequent TLVs.

(4.6)	*Umat*	*manusia*	***menghadapi***	***tantangan***	*penyelamatan*	***l**ingkungan*	*hidup.*
	worshiper	mankind	faces	challenge	rescue	environtment	life

'Humanity faces the challenge of environmental conservation.'

(4.7)	*Kondisi*	*itu*	***membuka***	***peluang***	*terjadinya*	*polarisasi*	*politik.*
	condition	that	opens	opportunity	happening	polarization	politics

'That situation creates an opportunity for political polarization to occur.

(4.8)	*Program*	*ini*	***mendapat***	***dukungan***	*dari*	*beberapa*	*instansi.*
	program	this	receives	support	from	several	institutions

'This program receives support from several institutions.'

(4.9)	*Ia*	*dianjurkan*	***menjalani***	***perawatan***	*medis*	*di*	*rumah sakit.*
	he/she	is advised	undergo	treatment	medical	at	hospital

'He/she is advised to undergo medical treatment at the hospital.'

(4.10)	*Saya*	***mengajukan***	***pertanyaan***	*ini*	*di*	*tahun*	*1938.*
	I	propose	question	this	in	year	1938

'I posed this question in 1938.'

The second sample set of LVCs, as in (4.6) to (4.10) continues to feature verbs that align with criteria for TLVs, exhibiting tendency toward aspectual neutrality and functioning primarily as grammatical scaffolds. Verb such as '*menghadapi*' (to face), '*membuka*' (to open), '*mendapat*' (to receive), '*menjalani*' (to undergo), and '*mengajukan*' (to pose/submit) show a consistent tendency toward [-] values in the aksionsart matrix, particularly in momentariness, frequentativeness, and intensiveness. For instance, in '*menghadapi tantangan penyelamatan lingkungan hidup*', the complex noun phrase—'*tantangan penyelamatan lingkungan hidup*'—encodes the situational durativity and conceptual weight, while the verb '*menghadapi*' merely licenses the predicative frame. Similarly, in '*mendapat dukungan*', the support is construed as a passive and stative notion, and the verb contributes do dynamicity. While '*menjalani perawatan medis*' might suggest [+durative] under certain readings, this is again attributable to the noun *perawatan*, not the verb itself. Even in '*mengajukan pertanyaan*', often used in dialogic or formal registers, the temporal dynamics are noun-driven. Overall, these verbs demonstrate limited inherent aktionsart values and reinforce their classification as TLVs. Their contribution to the aspectual profile of the construction remains minimal, affirming that temporal interpretation in the LVCs is primarily modulated by the noun element and surrounding context.

4.1.2 Suboptimal TLVs

While the notion *optimal* denotes a TLV's capacity to combine with a diverse range of nouns, indicating a high degree of syntactic and semantic flexibility, *suboptimal* signifies a more limited combinatorial scope, suggesting that the TLV exhibits greater selectivity in its noun pairings. This dichotomy in distribution behavior suggests the existence of an underlying "optimal type-pairing rate" that governs the compatibility between TLVs and nouns within LVCs. This rate, influenced by both lexical and grammatical factors, reflecting the inherent affinity between specific TLVs and certain noun classes.

Table 4.3: Sample list of the suboptimal TLVs.

No	TLV	Optimal Type-Pairing Rate	Sample LVCs	PMW
1.	*memicu*	4	*memicu kemarahan* 'get angry'	777.008
2.	*memulai*	4	*memulai usaha* 'makes an effort'	2935.702
3.	*menaruh*	4	*menaruh minat* 'pay attention'	499.806
4.	*mencari*	4	*mencari solusi* 'find a solution'	7377.704
5.	*menjalankan*	4	*menjalankan usaha* 'run a small business'	2201.990
6.	*menjelang*	4	*menjelang akhir* 'brings to an end'	2458.247
7.	*menyebabkan*	4	*menyebabkan ganguan* 'cause annoyance'	2302.846
8.	*menyediakan*	4	*menyediakan layanan* 'provide service'	4623.873
9.	*terjadi*	4	*terjadi kesalahpahaman* 'to misunderstand'	466.004
10.	*berjalan*	3	*berjalan kaki* 'take a walk'	12176.567

As a matter of fact, Table 4.3 presents a list of LVCs with limited pairing rate. Most of the TLVs in this table exhibit a suboptimal pairing rate, with the majority having a value of 4 and one having a value of 3. This suggests that these less frequent TLVs have a more restricted combinatorial potential compared to more frequent TLVs, potentially indicating their specialized semantic roles or contextual preferences.

(4.11) *Anda juga pasti sering* ***mencari solusi*** *dari berbagai masalah.*
you also definitely often search solution from various problems
'You must also often look for solutions to various problems.'

(4.12) *Dia* ***memulai usaha*** *ini pada tahun 1999.*
he/she started business this in year 1999
'He/She started this business in 1999.'

(4.13) *Satu kata yang salah ini* ***memicu kemarahan*** *massa.*
one word that wrong this triggered anger mass
'This one word triggered the anger of the masses.'

(4.14) *Para pemikir sudah lama* ***menaruh minat*** *pada olahraga ini.*
many thinker already long put interest on sport this
'Thinkers have long been interested in this sport.'

(4.15) *Mereka* ***menyediakan layanan*** *pencarian secara gratis.*
they provide service search in a manner free
'They provide search services for free.'

Particularly, sample in (4.11) to (4.15) feature a set of suboptimal TLVs, defined here by their limited co-occurrence with noun types—typically restricted to 3-4 pairings. Despite their classification as TLVs, these verbs, such as '*mencari*' (to search), '*memulai*' (to start), '*memicu*' (to trigger), '*menaruh*' (to place), and '*menyediakan*' (to provide) demonstrate varying degrees of aktionsart contribution but generally lean toward partial semantic bleaching. Most of these verbs tend toward [-frequentative], [-terative], and [-intensive], but may show context-sensitive [+momentary] or [+durative] values. For instance, *'memicu kemarahan'* (trigger anger) involves a punctual effect, suggesting [+momentary], while '*menyediakan layanan pencarian*' (provide search services) entails an extended action, indicating [+durative]. However, even in these cases, the temporal contour is heavily modulated by the noun component (*kemarahan, layanan*) rather than the verb itself. The limited type-pairing range of these verbs may arise from semantic opacity, which constraints their collocational spread despite functioning structurally as TLVs. Thus, while these verbs exhibit fewer co-occurrence patterns, they retain the key TLV characteristic: a tendency toward aktionsart underspecification, allowing the noun to drive the eventive and temporal interpretation of the LVC.

Moreover, Table 4.4 provides a complementary analysis of ten TLVs with suboptimal distribution within Indonesian LVCs. These verbs, while exhibiting the characteristics of TLVs, demonstrate limited combinatorial potential, suggesting specialized semantic roles or contextual preferences. The consistent optimal pairing rate of 3 across all ten TLVs emphasizes their limited combinatorial possibilities. This suggests that these verbs are highly selective in their noun pairings, potentially indicating their specialized semantic roles or contextual preferences.

Table 4.4: Extended-sample list of the suboptimal TLVs.

No	TLV	Optimal Type-Pairing Rate	Sample LVCs	PMW
1.	*memperoleh*	3	*memperoleh pendapatan* ‘do receiving’	732.004
2.	*memukul*	3	*memukul gong* ‘strike gongs’	316.529
3.	*menggelar*	3	*menggelar tahapan* ‘make a step’	6.379
4.	*menggunakan*	3	*menggunakan transportasi* ‘do transportation’	770.076
5.	*mengundang*	3	*mengundang kekaguman* ‘have admiration’	102.362
6.	*meninggalkan*	3	*meninggalkan goresan* ‘get scratched’	35.661
7.	*menjaga*	3	*menjaga kehormatan* ‘defend honor’	1030.954
8.	*menjatuhkan*	3	*menjatuhkan putusan* ‘give verdict’	447.068
9.	*menunjukkan*	3	*menunjukkan keangkuhan* ‘to brag’	6.128
10.	*meraih*	3	*meraih pencapaian* ‘to put upright on the legs’	89.152

Despite their restricted combinatorial potential, the sample LVCs demonstrate the semantic diversity of these TLVs. They encompass various domains, including receiving as in ‘*memperoleh pendapatan*’ (to receive income), physical actions as in ‘*memukul gong*’ (to strike gongs), sequential actions as in ‘*menggelar tahapan*’ (to make a step), transportation as in ‘*menggunakan transportasi*’ (to use transportation), admiration as in ‘*mengundang kekaguman*’ (to inspire admiration), physical markings as in ‘*meninggalkan goresan*’ (to leave a scratch), honor as in ‘*menjaga kehormatan*’ (to defend honor), legal decisions as in ‘*menjatuhkan putusan*’ (to give a verdict), boasting as in ‘*menunjukkan keangkuhan*’ (to show arrogance), and achievement as in ‘*meraih pencapaian*’ (to achieve achievements).

(4.16) *Manusia* ***meraih*** ***pencapaian-pencapaian*** *terhebatnya* *di* *era* *ini.*
human achieve achievements greatest-his in era this
‘Humas are achieving their greatest achievements in this era.’

(4.17) *Mahkamah Konstitusi RI* *dapat* ***menjatuhkan*** ***putusan*** *yang* *seadil-adilnya.*
constitutional court of Indonesia can drop decision which as-just-as-possible
‘The Constitutional Court of Indonesia can deliver the justest possible.’

(4.18) *Mereka* *sangat* ***menjaga*** ***kehormatan*** *dirinya.*
they very guard honor self-their
'They strongly uphold their honor.'

(4.19) *Atraksi* *itu* ***mengundang*** ***kekaguman*** *master* *karate* *dari* *Jepang* *tersebut.*
attraction that invite admiraton master karate from Japan that
'That attraction inspired the admiration of the karate master from Japan.'

(4.20) *Wajar* *kita* ***memperoleh*** ***pendapatan*** *aktif.*
natural we obtain income active
'It is reasonable for us to earn an active income.'

Particularly, the additional set of suboptimal TLVs, as in (4.16) to (4.20), further illustrates the nuanced role of aspectual neutrality in Indonesian LVCs. Verbs such as '*meraih*' (to attain), '*menjatuhkan*' (to deliver/to drop), '*menjaga*' (to guard), '*mengundang*' (to invite), and '*memperoleh*' (to obtain) each appear in constructions with constrained collocational range—typically combining with only three noun types in the dataset. Despite their surface lexical surface richness, these verbs exhibit a consistent tendency toward aktionsart underspecification, particularly in the dimensions of [-frequentative], [-iterative], and [-intensive]. For example, in '*meraih pencapaian*', the momentariness of the action is largely noun-driven, with '*meraih*' functioning as a transitional operator rather than a semantically loaded verb. Similarly, *'menjatuhkan putusan*' (deliver a verdict) may suggest [+momentary], yet the temporal resolution is embedded in *putusan,* not the verb itself. Even in *'menjaga kehormatan*', where durativity might be inferred, it is the enduring nature of *kehormatan* (honor) that licenses that reading. These patterns confirm that even suboptimal TLVs—despite their more selective noun compatibility—still align with the core TLV property: minimal contribution to the temporal contour of the event. Their limited pairing scope may result from semantic specificity, institutionalized usage, or restricted discourse domains, yet their aktionsart profiles affirms their role as grammatical enablers within LVC system.

4.2 Vague action verbs

In contradistinction to TLVs, VAVs within Indonesian LVCs manifest a subtly attenuated semantic weight and do not demonstrate a significant reliance on the accompanying nominal constituent for their semantic instantiation. While VAVs may convey a general sense of action or activity, their inherent aspectual specification is not appreciably more constrained, nor is it subject to substantial influence and modulation by the nominal element within the LVC. As elucidated by the theoretical underpinnings in Chapter 1, VAVs frequently manifest a composite profile of *positive* [+] and *negative* [-] values with respect to temporal properties, their specific interpretation remaining relatively independent of the semantic contribution of the noun. For instance, a VAV may evince [+ durative] but [- intensive] characteristics, contingent upon the specific nominal with which it combined. Thus, unlike TLVs that primarily serve grammatical functions, VAVs retain some semantic content and contribute more actively to the meaning of the construction to a certain degree.

4.2.1 Optimal VAVs

According to the analysis of VAVs, Table 4.5 provides a preview into the optimal VAVs within Indonesian LVCs. The consistent optimal pairing rate of 2 across all ten VAVs suggests a relatively restricted combinatorial range for these verbs compared to the TLVs. This might indicate that these frequent VAVs have more specialized semantic roles or contextual preferences compared to verbs with higher pairing rates.

Table 4.5: Sample list of the optimal VAVs.

No	VAV	Optimal Type-Pairing Rate	Sample LVCs	PMW
1.	*kehilangan*	2	*kehilangan akal* ‘to lose one’s head’	702.672
2.	*masuk*	2	*masuk akal* ‘to make sense’	16412.941
3.	*melahirkan*	2	*melahirkan gerutu* ‘give a grunt’	1,055
4.	*melalui*	2	*melalui mekanisme* ‘put something into operation’	3044.644
5.	*melemparkan*	2	*melemparkan senyuman* ‘to smile’	251.335
6.	*melihat*	2	*melihat keadaan* ‘monitor’	1064.957
7.	*memainkan*	2	*memainkan peran* ‘plays a role’	8085.299
8.	*membalas*	2	*membalas dendam* ‘taking a furious revenge’	2400.336
9.	*membawakan*	2	*membawakan tarian* ‘to dance’	199.551
10.	*membentuk*	2	*membentuk pemikiran* ‘do think’	23.607

Despite their limited pairing rate, the sample LVCs demonstrate the semantic diversity of these VAVs. They encompass various domains, including loss of control as in ‘*kehilangan akal*’ (to lose one’s head’), making sense as in ‘*masuk akal*’ (to make sense), physical expressions as in ‘*melahirkan gerutu*’ (to give a grunt), operational processes as in ‘*melalui mekanisme*’ (to put something into operation), facial expressions as in ‘*melemparkan senyuman*’ (to smile), observation as in ‘*melihat keadaan*’ (to monitor), role-playing as in ‘*memainkan peran*’ (to play a role), revenge as in '*membalas dendam*' as in (to take revenge), artistic performance as in '*membawakan tarian*' (to dance), and cognitive processes as in '*membentuk pemikiran*' (to think).

(4.21) *Pemerintah sudah* ***kehilangan akal.***
government already lost mind
'The government has lost its senses.'

(4.22) *Menurut saya ini* ***masuk akal.***
according me this enter sense
'In my opinion, this makes sense.'

(4.23) *Mereka senantiasa* ***melemparkan senyuman*** *yang mesra.*
they always throw smile which friendly
'They always give friendly smiles.'

(4.24) *Organisme juga* ***memainkan peran*** *dalam proses itu.*
organisms also play role in process that
'Organisms also play a role in that process.'

(4.25) *Anusapati ingin* ***membalas dendam*** *pada Ken Angrok.*
Anusapati want reply revenge on Ken Angrok
'Anusapati wants to take revenge on Ken Angrok.'

Furthermore, the present sample as in (4.20) to (4.25), features a series of LVCs whose verb elements align with the characteristics of VAVs, as defined in the aktionsart framework. Unlike TLVs, which are semantically bleached, VAVs contribute a modest yet discernible degree of aspectual meaning to the construction. Verb such as '*kehilangan*', '*masuk*', '*melemparkan*', '*memainkan*', and '*membalas*' display a range of [+] and [-] values across the five temporal features. For example, '*kehilangan akal*' (lose one's senses) suggests a [+momentary] change-of-state reading, while '*masuk akal*' (make sense) may lean toward [-durative] and [-intensive], reflecting its stative epistemic interpretation. Thereafter, '*melemparkan senyuman*' [+durative] and potentially [+iterative] if interpreted in a habitual context, though these readings depend heavily on discourse context. Similarly, '*memainkan peran*' (play a role) and '*membalas dendam*' (take revenge) suggest [+durative] and [+intensive], yet these properties are shaped by the semantics of *peran* and *dendam*, respectively. What distinguishes these verbs from TLVs is their partial aktionsart specification, which allows them to contribute generalized actional contours without fully determining the temporal semantics of the LVC. This intermediate position supports their classifications as VAVs: verbs that mediate between grammatical support and semantic contribution, and whose temporal features are sensitive to the interpretative role of the noun complement.

Table 4.6 offers a focused secondary look at ten VAVs that appear optimal within Indonesian LVCs. The consistent optimal pairing rate of 2 across all ten VAVs suggests a circumscribed combinatorial range for these verbs. This might indicate that these VAVs have even more specialized semantic roles or contextual preferences compared to Table 4.5.

Table 4.6: Extended-sample list of the optimal VAVs.

No	VAV	Optimal Type-Pairing Rate	Sample LVCs	PMW
1.	*membubuhkan*	2	*membubuhkan tanda* 'give a sign'	1071.889
2.	*memunculkan*	2	*memunculkan desakan* 'give a prod'	5.927
3.	*menarik*	2	*menarik diri* 'to pull out'	1481.739
4.	*menawarkan*	2	*menawarkan bantuan* 'give assistance'	634.665
5.	*menciptakan*	2	*menciptakan kemelut* 'make noise'	23.858
6.	*mencium*	2	*mencium aroma* 'have a smell perceived'	821.659
7.	*mencuri*	2	*mencuri hati* 'steal someone's heart'	326.122
8.	*mengangkat*	2	*mengangkat topik* 'to bring up a topic'	405.732
9.	*mengenai*	2	*mengenai keberadaan* 'do existence'	742.401
10.	*mengikat*	2	*mengikat hati* 'to make someone fall in love'	156.908

Despite their limited pairing rate, the sample LVCs demonstrate the semantic diversity of these VAVs. They encompass various domains, including symbolic actions as in '*membubuhkan tanda*' (to give a sign), persuasion or pressure as in '*memunculkan desakan*' (to give a prod), withdrawal or retreat as in '*menarik diri*' (to pull out), offering assistance as in '*menawarkan bantuan*' (to offer assistance), creating disturbances as in '*menciptakan kemelut*' (to create chaos), olfactory perception as in '*mencium aroma*' (to smell a scent), emotional impact as in '*mencuri hati*' (to steal someone's heart), introducing topics as in '*mengangkat topik*' (to bring up a topic), addressing existence as in '*mengenai keberadaan*' (to address existence), and romantic pursuit as in '*mengikat hati*' (to make someone fall in love).

(4.26) *Mereka* *kemudian* ***membubuhkan*** ***tanda*** *tangan* *di* *spanduk.*
they then affixed sign hand at banner
'They then put their signatures on the banner.'

(4.27) *Peneliti* *dapat* ***menarik*** ***diri*** *setelah* *durasi* *penelitian* *selesai.*
researcher can withdraw self after duration research finished
'The researcher can withdraw after the research period is completed.'

(4.28) *Kami* **menawarkan** **bantuan** *kepada* *anda.*
we offer help to you
'We offer you help.'

(4.29 *Dengan* *gaya* *yang* **mencuri** **hati** *dia* *menjadi* *petugas* *operator* *telepon.*
with style that steal heart she became officer operator telephone
'With a captivating style, she became a telephone operator.'

(4.30) *Rubrik* *ini* **mengangkat** **topik** *menarik* *yang* *ditanyakan* *oleh* *pendengar.*
column this lift topic interesting that asked by listener
'This column raises an interesting topic that was asked by a listener.'

Lastly, the subsequent sample of LVCs, in (4.25) to (4.30), built with VAVs, while not fully semantic rich, contribute a partial and context-sensitive aspectual layer to the construction. Verbs like '*membubuhkan*' (to affix) and '*menarik*' (to pull/withdraw) indicate [+momentary] and potentially [+intensive] values when paired with nouns like '*tanda tangan*' or reflexsive pronouns '*diri*'. These constructions imply brief, agentive acts embedded within bounded events. By contrast '*menawarkan bantuan*' (to offer help) evokes a [+durative] frame with optional [+iterative] interpretation, especially in institutional or service-oriented discourse. Thereafter, '*mencuri hati*' (to steal the heart) exhibits metaphorical transfer, suggesting [+intensive] and contextually [+iterative] readings, particularly in affective or persuasive settings. Also, '*mengangkat topik*' (to raise a topic) adds eventive dynamicity, potentially [+frequentative] depending on its discursive role. Collectively, these verbs do not fulfill the complete aktionsart profile of lexical verbs, yet they surpass TLVs in temporal contribution, especially where nominal elements alone cannot encode event structure. Their partial semantic content enables them to occupy a middle space between syntactic facilitation and event specification. The analysis affirms that VAVs operate with gradient aktionsart features, modulated by semantic contour of their nominal complement.

4.2.2 Suboptimal VAVs

In addition to the optimal VAVs, there is also the opposite tendency. Table 4.7 offers a focused exploration of ten VAVs that appear suboptimal within Indonesian LVCs. The predominant optimal pairing rate of 2 across most VAVs in the table suggests a restricted combinatorial range for these verbs. This might indicate that these less frequent VAVs have even more specialized semantic roles or contextual preferences compared to the more frequent VAVs. The presence of two verbs with a pairing rate of 1 further emphasizes this limited combinatorial potential.

Table 4.7: Sample list of the suboptimal VAVs.

No	VAV	Optimal Type-Pairing Rate	Sample LVCs	PMW
1.	*mengikuti*	2	*mengikuti arus* ‘to go with the flow’	480.067
2.	*menginjak*	2	*menginjak rem* ‘put a brake on’	303.721
3.	*mengulurkan*	2	*mengulurkan tangan* ‘give your hand’	1197.204
4.	*menjual*	2	*menjual diri* ‘to sell oneself’	232.650
5.	*menyalurkan*	2	*menyalurkan bantuan* ‘to provide help’	1712.330
6.	*menyatakan*	2	*menyatakan persetujuan* ‘do agree’	54.647
7.	*pergi*	2	*pergi rekreasi* ‘go on trip’	9.593
8.	*turun*	1	*turun tangan* ‘to take action’	4340.645
9.	*tertangkap*	1	*tertangkap basah* ‘to be caught red-handed’	1891.639
10.	*tidur*	1	*tidur siang* ‘take a nap’	1749.196

Despite their limited pairing rate, the sample LVCs reveal the semantic variety of these VAVs. They involve several domains, including following a trend as in ‘*mengikuti arus*’ (to go with the flow), stopping or hindering as in ‘*menginjak rem*’ (to put a brake on), offering help ‘*mengulurkan tangan*’ (to give your hand), self-sacrifice as in ‘*menjual diri*’ (to sell oneself), providing assistance as in ‘*menyalurkan bantuan*’ (to provide help), expressing agreement as in ‘*menyatakan persetujuan*’ (to agree), leisure activities as in ‘*pergi rekreasi*’ (to go on a trip), taking action ‘*turun tangan*’ (to take action), being caught in the act as in ‘*tertangkap basah*’ (to be caught red-handed), and taking a rest as in ‘*tidur siang*’ (to take a nap).

(4.31) *Mereka hanya* ***mengikuti arus*** *tren apa yang populer di internet.*
they only follow current trend what is popular on internet
‘They only follow the current of trends that are popular on the internet.’

(4.32) *Para petugas* ***mengulurkan tangan*** *ingin membantu.*
several officer extend hand want help
‘The officer extended a hand wanting to help.’

(4.33) *Aku tidak* ***menjual diri.***
I not sell self
‘I don’t sell myself.’

(4.34) *Pemerintah* ***turun tangan*** *dan menyatakan penolakan.*
government descend hand and declare rejection
‘The government descended hand and declared a rejection.’

(4.35) *Mereka* ***tertangkap*** ***basah*** *tengah* *bermain* *judi.*
they caught wet middle playing gambling
'They were caught red-handed in the middle of playing gambling.'

Specifically, the subsequent set of samples in (4.30) to (4.35) highlights the behavior of suboptimal VAVs within LVCs system. Verbs such as '*mengikuti*' (to follow), '*mengulurkan*' (to extend), '*menjual*' (to sell), '*turun*' (to descend), and '*tertangkap*' (to be caught) illustrate this limited productivity. From the aktionsart perspective, '*mengikuti arus*' (follow the current of trends) is [+durative] and [+frequentative], particularly in habitual online behavior contexts. Thereafter, '*mengulurkan tangan*' (extend a hand) suggests [+momentary] and [+intensive], reflecting a punctual, agentive gesture with emotive weight. In addition, '*menjual diri*' (sell oneself) may carry [+intensive] and [+iterative] readings in socio-moral discourse, though often interpreted metaphorically. Also, '*turun tangan*' (descend hand) encodes a metaphorical intervention act, semantically [+momentary] and [+intensive] but discursively marked. Finally, '*tertangkap basah*' (caught red-handed) is passive, stative construction implying [+momentary] and [+intensive] surprise of transgression. While these verbs contribute more aspectual content than TLVs, their narrow type-pairing rate indicates semantic specificity, idiomatic rigidity, or genre-dependence, which limits their generalization. This reinforces their classification as VAVs with limited systemic reach, revealing a trade-off between semantic salience and combinatiorial openness in the LVC paradigm.

Table 4.8: Extended-sample list of the suboptimal VAVs.

No	VAV	Optimal Type-Pairing Rate	Sample LVCs	PMW
1.	*tutup*	1	*tutup buku 'to end something'*	513.267
2.	*ada*	1	*ada ekornya 'there's more to it'*	2.662
3.	*adu*	1	*adu domba 'to pit against each other'*	1320.411
4.	*angkat*	1	*angkat topi 'to take one's hat off (to someone)'*	426.576
5.	*tunjuk*	1	*tunjuk hidung 'to point a finger (at someone)'*	66.500
6.	*tertutup*	1	*tertutup pikirannya 'to be closed-minded'*	0.603
7.	*tertusuk*	1	*tertusuk hatinya 'to be deeply hurt'*	4.621
8.	*terserang*	1	*terserang flu 'get the flu'*	664.449
9.	*tersentuh*	1	*tersentuh hatinya 'to be touched'*	101.257
10.	*terpikat*	1	*terpikat hatinya 'to be captivated'*	11.753

Furthermore, the analysis also reveals a list of constricted distribution VAVs. Table 4.8 provides a focused exploration of ten VAVs with constricted distribution within Indonesian LVCs. These verbs, while exhibiting some characteristics of light verbs, demonstrate a limited range of noun pairings, suggesting specialized semantic roles or contextual preferences. The consistent optimal pairing rate of 1 across all ten VAVs emphasizes their extremely limited combinatorial possibilities. This suggests that these verbs are highly selective in their noun pairings, potentially indicating their highly specialized semantic roles or contextual preferences.

Despite their restricted combinatorial potential, the sample LVCs demonstrate the semantic diversity of these VAVs. They encompass various domains, including concluding or ending something as in '*tutup buku*' (to end something), having additional information or consequences as in '*ada ekornya*' (there's more to it), instigating conflict as in '*adu domba*' (to pit against each other), showing respect as in '*angkat topi*' (to take one's hat off to someone), expressing disapproval as in '*tunjuk hidung*' (to point a finger at someone), having a closed mind as in '*tertutup pikirannya*' (to be closed-minded), experiencing deep emotional hurt as in '*tertusuk hatinya*' (to be deeply hurt), contracting an illness as in '*terserang flu*' (to get the flu), being emotionally touched as in '*tersentuh hatinya*' (to be touched), and being captivated or enchanted as in '*terpikat hatinya*' (to be captivated).

(4.36) *Ini* *adalah* ***tutup*** ***buku*** bagi *logika* *formal.*
this is close book for logic formal
'This is the end of the discussion for formal logic.'

(4.37) *Ia* *melakukan* *praktik* ***adu*** ***domba*** *itu* *berdasarkan* *pendapatnya.*
he/she do practice pit sheep that based on his/her opinion
'He/She carried out the practice for pitting people against each other based on his/her opinion.'

(4.38) *Saya* ***angkat*** ***topi*** *buat* *Nezar.*
I lift hat for Nezar
'I take my hat off to Nezar.'

(4.39) *Direktur* *langsung* ***tunjuk*** ***hidung*** *kepada* *A.*
Director directly point finger to A
'The director directly pointed his finger at A.'

(4.40) *Sang* *Prabu* *Kertajasa* ***terpikat*** ***hatinya*** *oleh* *kecantikan* *sang* *puteri* *ini.*
the king Kertajasa captivated his heart by beauty the princess this.
'King Kertajasa was captivated by the beauty of this princess.'

Moreover, the final subset of LVCs as in (4.35) to (4.40) exemplifies suboptimal VAVs that retain partial semantic value yet occur exclusively with a single noun pairing, indicating high idiomatic specificity and low combinatorial freedom. Examples such as '*tutup buku*' (close the book), '*adu domba*' (carry out a divise practive), '*angkat topi*' (lift one's hat). '*tunjuk* *hidung*' (point a finger), and '*terpikat hati*' (be captivated in heart) are all structurally LVCs but semantically bound to fixed collocations. According to the aktionsart matrix, these verbs tend to display selective [+] values across temporal properties. For instance, '*tutup buku*' suggests [+momentary] and [+intentsive] when used to signify abrupt termination of discourse. Additionally, '*adu domba*' leans toward [+durative] and [+frequentative], reflecting a sustained and potentially repetitive behavioral critique. Thereafter, '*angkat topi*' evokes a ceremonial act of respect, marked as [+momentary] and [+intensive]. Next, '*tunjuk hidung*' implies [+momentary] accusation, and '*terpikat hati*' is best classified as stative, yet [+intensive], given its emotive salience. While these verbs contribute more semantic content than TLVs, their extremely narrow pairing rate reflects lexical entrenchment within culturally or pragmatically fixed expressions. These VAVs thus function at the intersection of aktionsart selectivity and idiomaticity, showing that semantic richness can co-exist with syntactic rigidity in the lower-frequency margins of the Indonesian LVC system.

4.3 Assessment of the verb element within LVCs

The subsequent analysis leverages the capabilities of several machine learning algorithms to yield a more profound understanding of the verbal element within Indonesian LVCs. Two distinct computational operations are employed, each designed to extract unique insights from the intricate dataset of verbs embedded within these constructions. First, the implementation of a NB classifier will enable us to predict the categorical affiliation of verbs based on their observed features, contributing to a probabilistic model of verb selection within LVCs. Second, the utilization of regression random forests will allow us to establish predictive relationships between the verbs' linguistic features and their tendency of frequency, offering a quantitative perspective on the factors influencing their selection and usage within LVCs. Through the synergistic application of these machine learning techniques, we anticipate generating a multifaceted portrait of the verbal landscape within Indonesian LVCs, thereby enriching our empirical and theoretical understanding of their formation and function.

4.3.1 Assessment based on Naïve Bayes classifier

In the realm of Indonesian LVCs, the NB classifier serves as a tool for discerning the relationship between verb and its optimal pairing rate with nouns. By leveraging the principles of Bayesian probability, this algorithm allows us to predict the likelihood of a verb being classified as either a TLV or VAV based on its co-occurrence patterns with specific nouns. This prediction is achieved by analyzing the frequency with different verbs and nouns appear together in LVCs within a given corpus. The NB classifier then calculates the conditional probability of a verb belonging to a particular class (TLV or VAV) given its observed pairings with various nouns. This process enables us to identify those verbs that exhibit a strong tendency to combine with nouns in LVCs, thereby establishing a hierarchy of optimal pairing rates (*see* Table 4.9). The number following each verb in the table indicates its optimal number of the distinctive combination with noun elements. In detail, the complete list of verb forms is available in Appendix B.

Table 4.9: Sample list of the verb pairing-rate.

Verb element	Pairing rate	Verb element	Pairing rate
melakukan 'to do/to carry out'	141	*mengajukan* 'to propose/ to submit'	5
memberikan 'to give'	120	*mengeluarkan* 'to release/ to issue'	5
membuat 'to make'	53	*melepaskan* 'to let go'	4
mengalami 'to experience'	34	*membutuhkan* 'to need'	4
memberi 'to give'	31	*memicu* 'to trigger'	4
mengambil 'to take'	21	*memulai* 'to start'	4
menjadi 'to become'	18	*menaruh* 'to put/ to place'	4
memiliki 'to have'	16	*mencari* 'to search'	4
menyampaikan 'to deliver/ to convey'	16	*menjalankan* 'to run/ to operate'	4
mengadakan 'to hold/ to organize'	12	*menjelang* 'to approach/ in the lead-up to'	4
menimbulkan 'to cause'	12	*mengajukan* 'to propose/ to submit'	5
membawa 'to bring'	11	*mengeluarkan* 'to release/ to issue'	5
merasa 'to feel'	10	*melepaskan* 'to let go'	4
melontarkan 'to hurl/ to throw'	9	*membutuhkan* 'to need'	4
memasang 'to install/ to replace'	8	*memicu* 'to trigger'	4
menderita 'to suffer'	8	*memulai* 'to start'	4
melaksanakan 'to implement'	7	*menaruh* 'to put/ to place'	4
mendapatkan 'to obtain/ to receive'	7	*mencari* 'to search'	4
menghadapi 'to face'	7	*menjalankan* 'to run/ to operate'	4
melayangkan 'to launch/ to send forth'	6	*menjelang* 'to approach/ in the lead-up to'	4
membuka 'to open'	6	*mengajukan* 'to propose/ to submit'	5
mendapat 'to get'	6	*mengeluarkan* 'to release/ to issue'	5
menjalani 'to undergo'	6	*melepaskan* 'to let go'	4
merasakan 'to sense/ to feel'	6	*membutuhkan* 'to need'	4
membangun 'to build'	5	*memicu* 'to trigger'	4

The NB classifier, in particular, proves effective in capturing the co-occurrence patterns between light verbs and nouns, estimating the optimal type-pairing rate. It serves as an empirical lens through which we can discern the affinity of specific light verbs for certain classes of nouns. Accordingly, this analysis employs a NB classifier to investigate the collocational tendencies of verbs within Indonesian LVCs. Specifically, the aim seeks to quantify the propensity of verbs to co-occur with nouns within LVC framework, drawing upon data extracted from four corpora (ILCC, SLIC, IDC, and IWC). The classifier provides a probabilistic output, predicting the likelihood of a given verb taking a noun as it collocates within an LVC. While inherently a binary classification task (verb with noun vs. verb without noun), the output offers a nuanced understanding by providing a probability score, rather than a simple yes/no decision. Specifically, the input feature for the NB classifier is the categorization of verbs within Indonesian LVCs into two distinct classes: TLVs and VAVs. This classification, established through linguistic pre-analysis or prior study-section (*see* Chapter 2), serves as the basis for training the classifier to recognize the collocational behavior associated with each verb type.

Table 4.10: The performance of ten most productive true light verb.

TLV	NB	DTree	SVM	Logistic	Avg.
melakukan	**0.60**	0.60	0.30	0.15	0.4125
memberikan	**0.85**	0.85	0.43	0.21	0.585
membuat	-0.19	-0.19	-0.1	-0.05	-0.1325
mengalami	**0.90**	0.90	0.45	0.22	0.6175
memberi	-0.05	-0.05	-0.03	-0.01	-0.035
mengambil	**0.94**	0.94	0.47	0.24	0.6475
menjadi	-0.11	-0.11	-0.05	-0.03	-0.075
memiliki	**0.30**	0.30	0.15	0.07	0.205
menyampaikan	-2.58	-2.58	-1.29	-0.64	-1.7725
mengadakan	-0.66	-0.66	-0.33	-0.17	-0.455

Table 4.10 delves into the performance of the ten most frequently occurring TLVs in Indonesian LVCs, employing a suite of machine learning classifiers to predict the optimal pairing rate between these verbs and their corresponding nouns. The table summarizes the F1 score for ten frequently occurring TLVs, evaluated across four machine learning classification models: NB, DTree (Decision Tree), SVM (Support Vector Machine (SVM), and Logistic Regression. The table's primary focus is to assess the effectiveness of the NB classifier in capturing optimal type-pairings between TLVs and noun complements within LVCs. The F1 score, which balances precision and recall, serves as an indicator of how well each model predicts productive pairings. While average F1 score across models are included, particular emphasis is placed on the NB results due to its probabilistic framework's sensitivity to co-occurrence patterns. High NB scores, normal score range from 0 to 1, reflect robust and consistent type-pairing behavior. In contrast, verbs with negative scores across models are

suggesting irregular pairing tendencies. These patterns support a computational approach to understanding LVC structure through distributional learnability.

A cursory glance at Table 4.10 reveals a pattern. The NB classifier consistently outperforms the other three algorithms across almost all light verbs. This observation underscores the efficacy of the NB approach in capturing the underlying statistical dependencies between light verbs and their noun complements. For instance, the light verb '*melakukan*' (to do/perform) exhibits an F-1 score of 0.6 with NB, significantly higher than the 0.3 achieved by SVM and the 0.15 by Logistic Regression. This superior performance suggests that the co-occurrence patterns between '*melakukan*' and its associated nouns are well-captured by the conditional independence assumption of NB. Similarly, the light verb '*memberikan*' (to give/provide) achieves an F-1 score of 0.85 with NB, again surpassing the other classifiers. This result highlights the strong association between '*memberikan*' and specific semantic classes of nouns, such as those denoting abstract concepts like '*perhatian*' (attention) or '*informasi*' (information). However, not all light verbs exhibit such clear-cut distinctions. The light verb '*membuat*' (to make/create) shows a relatively lower F-1 score of -0.19 across all classifiers. This finding suggests a more nuanced relationship between '*membuat*' and its noun complements, potentially indicating a wider range of semantic possibilities or a higher degree of ambiguity in its usage.

Table 4.11: The performance of ten most productive vague action verb.

VAV	NB	DTree	SVM	Logistic	Avg.
kehilangan	**0.35**	0.35	0.17	0.09	0.24
masuk	**0.50**	0.50	0.25	0.13	0.345
melahirkan	**0.66**	0.66	0.33	0.17	0.455
melalui	**0.78**	0.78	0.39	0.19	0.535
melemparkan	-1.18	-1.18	-0.59	-0.29	-0.81
melihat	**0.01**	0.01	0	0	0.005
memainkan	**1.17**	1.17	0.59	0.29	0.805
membalas	**0.80**	0.80	0.40	0.20	0.55
membawakan	-1.10	-1.10	-0.55	-0.28	-0.7575
membentuk	-1.99	-1.99	-1	-0.5	-1.37

Furthermore, Table 4.11 focuses on the VAVs within Indonesian LVCs, specifically the ten most frequently occurring ones. These verbs, characterized by their relatively general semantic content, pose a unique challenge in understanding the factors that influence their co-occurrence with specific nouns. The table presents the results of a machine learning-based analysis aimed at predicting the optimal pairing rate between these vague action verbs and their corresponding nouns. A close examination of Table 4.11 reveals a consistent trend. The NB classifier generally outperforms the other three algorithms across the majority of vague action verbs. This observation underscores the effectiveness of the NB approach in capturing the subtle statistical relationships between these verbs and their noun complements.

For instance, the vague action verb '*kehilangan*' (to lose) achieves an F-1 score of 0.35 with NB, surpassing the 0.17 obtained by SVM and the 0.09 by Logistic Regression. This superior performance suggests that the NB classifier, despite its simplifying assumptions, effectively captures the co-occurrence patterns between '*kehilangan*' and its associated nouns, such as '*muka*' (face) or '*akal*' (mind or sense). Similarly, the verb '*masuk*' (to enter) exhibits an F-1 score of 0.5 with NB, again outperforming the other classifiers. This result highlights the strong association between '*masuk*' and specific semantic categories of nouns, such as those denoting physical situations (*angin* 'wind') or abstract states (*akal* – 'mind or sense'). However, the performance varies across different verbs. The verb '*melemparkan*' (to throw) shows a negative F-1 score of -1.18 across all classifiers. This finding suggests a more complex relationship between '*melemparkan*' and its noun complements, potentially indicating a wider range of semantic possibilities or a greater degree of ambiguity in its usage within LVCs.

4.3.2 Assessment based on Random Forests classification model

To further the NB analysis presented in the preceding section, Regression Random Forests are employed to gauge the *importance* of both quantitative and qualitative variables pertinent to this investigation. The purpose of this test is to ascertain which variables exert a significant influence on the selection of verbal elements, specifically TLVs or VAVs, within Indonesian LVCs. The quantitative variables utilized are based on the frequency of occurrence of the entire dataset within each corpus analyzed. Four variables fall under this category: the frequency of LVCs in ILCC, SLIC, IDC, and IWC. Meanwhile, the qualitative variables are derived from theoretical linguistic principles, structured into six parameters: M-1, M-2, S-1, S-2, Sx-1, and Sx-2.

In detail, this analysis utilizes regression random forests to explore the factor influencing the classification of verbs within Indonesian LVCs. The analysis aims to address two primary questions: (a) to what extent does the quantitative variable of verb frequency, as observed across four selected corpora, predict the classification of a verb as either TLV or a VAV within Indonesian LVCs; and (b) to what extent does qualitative variable of grammatical distribution, represented by the verb's occurrence across six grammatical conditions or relations, predict the classification of a verb as either a TLV or a VAV within Indonesian LVCs? The regression random forest model provides both predictive output (estimating the probability of a verb belonging to each class) and explanatory output (identifying the relative importance of the quantitative and qualitative variables in determining verb classification). This allows for a nuanced understanding of the interplay between verb frequency, grammatical context, and their impact on the semantic role of verbs within LVCs.

Table 4.12: Corpora variable-importance for predicting the status of true light or vague action verbs.

Variables	Va	Tl	Mean decrease accuracy
ILCC	4.127	8.040	**13.040**
SLIC	-6.855	7.585	1.822
IDC	7.439	0.386	6.897
IWC	6.551	-2.080	4.457

First, regarding importance of quantitative variable, Table 4.12 presents the results of a variable importance analysis derived from a Random Forest regression model. In this context, the goal appears to be predicting whether a verb within an Indonesian LVCs's functions as a true light verb ('Tl') or a vague action verb ('Va'). The table highlights the relative importance of four linguistic corpora (ILCC, SLIC, IDC, and IWC) in making this prediction. There are several key observations, for ILCC, this corpus emerges as the most influential predictor, with the highest mean decrease in accuracy (13.040). Its substantial impact on both 'Va' (4.127) and 'Tl' (8.040) suggests that the linguistic patterns and verb usage within ILCC are highly informative in differentiating between true light verbs and vague action verbs within LVCs. For IDC, the IDC corpus also demonstrates considerable importance, with a mean decrease in accuracy of 6.897. While its impact on predicting vague action verbs ('Va': 7.439) is notable, its influence on predicting true light verbs ('Tl': 0.386) is minimal. This suggests that IDC might be particularly rich in features associated with vague action verbs. With a mean decrease in accuracy of 4.457, IWC holds moderate importance in the model. Interestingly, its impact is skewed towards predicting true light verbs ('Tl': 6.551), with a negative value for 'Va' (-2.080). This negative value could indicate that permuting IWC actually improves the model's prediction of vague action verbs, suggesting a complex relationship between this corpus and the verb classes. Lastly, for SLIC, this corpus exhibits the lowest overall importance (1.822). While it contributes slightly to predicting both verb types, its impact is relatively minor compared to the other corpora as illustrated in Figure 4.3.

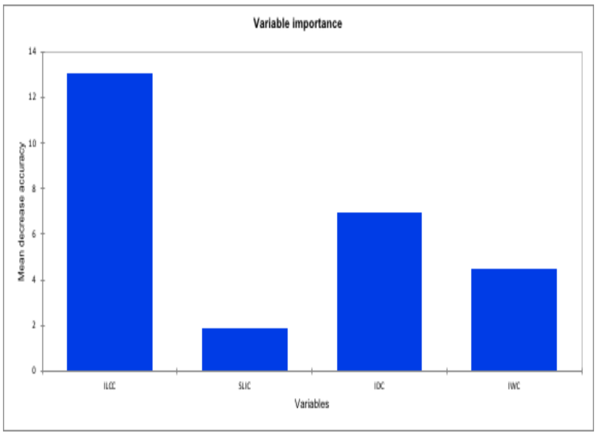


Figure 4.3: Variable importance of the quantitative aspects for predicting the status of true light or vague action verbs.

Therefore, several implications can be deduced from this observation. As of corpus specificity, the varying importance of the corpora suggests that the linguistic features and contexts captured in each corpus differ in their relevance to distinguishing between true light verbs and vague action verbs in Indonesian LVCs. Accordingly, ILCC's prominence is a notable one. The high importance of ILCC underscores its value as a resource for understanding the subtle distinctions between these verb types. Additionally, the strong association of IDC with vague action verbs suggests that this corpus may contain a wealth of examples or linguistic contexts that shed light on the nature and usage of these verbs within LVCs. Lastly, IWC and TLVs: The positive impact of IWC on predicting true light verbs and its negative impact on predicting vague action verbs points to a potential contrast in the linguistic features or contexts represented in this corpus.

Table 4.13: Theoretical linguistics variable-importance for predicting the status of true light or vague action verbs.

Variables	Tl	Va	Mean decrease accuracy
M-1	0.176	12.121	8.736
M-2	0.036	10.200	7.398
S-1	11.516	14.444	18.327
S-2	12.750	20.926	**23.801**
Sx-1	19.476	5.354	21.683
Sx-2	14.866	-4.296	11.269

In addition, regarding importance of qualitative variable, Table 4.13 presents the results of a variable importance analysis, originated from a Random Forest regression model, within the context of distinguishing between true light verbs (TL) and vague action verbs (VA) in Indonesian LVCs. This analysis specifically focuses on theoretical linguistic principles related to morphology, semantics, and syntax, offering insights into the linguistic factors that contribute most significantly to this distinction. Several key observations can be formulated. First, Semantic Principles Dominate. The semantic principles, S-1 (scale of synonymity) and S-2 (prototypicality), emerge as the most influential predictors, with the highest mean decrease in accuracy values (18.327 and 23.801, respectively). This suggests that the semantic relationships within LVCs, particularly the presence or absence of synonymous counterparts for the verb and the degree to which the noun carries the core propositional content, are crucial in distinguishing true light verbs from vague action verbs.

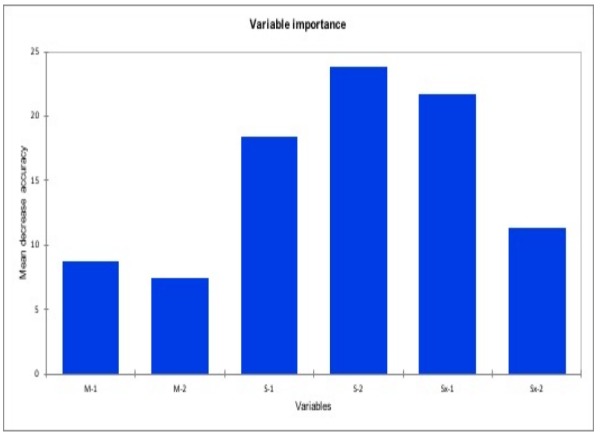


Figure 4.4: Variable importance of the qualitative aspects for predicting the status of true light or vague action verbs.

Second, Syntactic Principle of Transitivity (Sx-1). This principle also demonstrates high importance, with a mean decrease in accuracy of 21.683. This implies that the transitivity of the verb within the LVC (i.e., whether it takes a direct object) is a significant factor in determining its status as a true light verb or a vague action verb. Third, Morphological Principles (M-1 and M-2). The morphological principles, M-1 (“base” vs. “affixed”) and M-2 (“abstract” vs. “concrete”), exhibit lower importance compared to the semantic and syntactic principles. While they contribute to the model’s predictive power, their influence is less pronounced, as demonstrated in Figure 4.4. Fourth, Syntactic Principle of Valency (Sx-2). Interestingly, this principle shows a negative mean decrease in accuracy for vague action verbs (-4.296). This suggests that permuting this principle actually improves the model’s ability to predict vague action verbs. This counterintuitive result warrants further investigation and may indicate complex interactions between valency and other linguistic factors in determining verb status within LVCs.

Therefore, based on the evaluation of qualitative variable aforementioned, there are several implications. Regarding the semantics as a key differentiator. The prominence of the semantic principles underscores the importance of meaning and conceptual relationships in distinguishing true light verbs from vague action verbs. This suggests that the semantic contribution of the noun within the LVC and the availability of synonymous alternatives for the verb play a crucial role in determining its function. As of syntactic considerations, the significance of the transitivity principle highlights the relevance of syntactic structure and argument realization in classifying verbs within LVCs. Also, morphological nuances, while less influential than semantic and syntactic factors, morphological properties of the verb, such as its base form or affixation, and its degree of abstractness or concreteness, still contribute to the identification of true light verbs and vague action verbs. Lastly, concerning valency and complexity, the unexpected negative impact of the valency principle on predicting vague action verbs suggests a complex interplay between the number of arguments a verb takes and its status within LVCs.

4.4 Discussion

This discussion is primarily dedicated to a critical examination of the verbal element within Indonesian LVCs. The central theoretical framework underpinning this analysis is the exploration of the intricate (inter)dependency between the verb and noun components within these constructions. To achieve this, we will draw upon selected results from pertinent evaluation metrics, providing forceful empirical support for our theoretical investigations. In particular, we will focus on those metrics that shed light on the semantic and syntactic interplay between verbs and nouns in LVCs, examining how the choice of verb influences the interpretation and function of the noun, and vice versa. Specifically, there are several notable aspects, i.e. qualitative and quantitative variables for verb elements predictions (§4.4.1), type of nouns as the predictor of verb element (§4.4.2), and classes within stative and eventive noun as the predictor for verb-element selection (§4.4.3). By carefully analyzing these interactions, we aim to gain a deeper understanding of the complex dynamics at play within Indonesian LVCs and their implications for linguistic theory and language processing.

4.4.1 Qualitative and quantitative variables for verb elements predictions

As the subsequent analysis and discussion, Table 4.14 presents the results of a mix-variable importance analysis originated from a Random Forest model. This analysis adopts a more holistic approach by considering a mix of quantitative and qualitative variables to assess their predictive power in distinguishing between true light verbs (TL) and vague action verbs (VA) within Indonesian LVCs. This multifaceted perspective offers a richer understanding of the factors that contribute to verb classification in these constructions. As the first key observation, `TYPE_NOUN` as the Dominant Predictor. The `TYPE_NOUN` variable emerges as the most influential predictor, with the highest mean decrease in accuracy (20.879). Its substantial impact on both 'tl' (20.313) and 'va' (9.394) classes underscores its crucial role in differentiating between true light verbs and vague action verbs. This suggests that the semantic class of the noun within the LVC is a strong indicator of the verb type, as illustrated in Figure 4.5, highlighting the close interplay between the verb and noun in these constructions.

Table 4.14: Mix variable-importance for predicting the status of true light or vague action verbs.

Variables	tl	va	Mean decrease accuracy
TOTAL FREQ	12.343	0.023	13.947
RANK	16.170	-1.331	16.063
TYPE NOUN	20.313	9.394	**20.879**

Second, as of `RANK` with moderate importance, the `RANK` variable demonstrates moderate predictive power, with a mean decrease in accuracy of 16.063. While its impact on predicting true light verbs (tl: 16.170) is notable, its influence on predicting vague action verbs is negative (-1.331). This negative value suggests that permuting the rank of LVCs actually improves the model's ability to identify vague action verbs. This counter-intuitive result might indicate that the relative frequency or ranking of LVCs is less informative for classifying vague action verbs, potentially due to their greater flexibility in combining with various nouns and appearing in diverse contexts. Third, `TOTAL_FREQ` with mixed impact. The `TOTAL_FREQ` variable exhibits a mixed impact, with a high mean decrease in accuracy (13.947) but a negligible effect on predicting vague action verbs (va: 0.023). This suggests that the overall frequency of the LVC is more informative for classifying true light verbs, potentially reflecting their tendency to occur in more frequent and conventionalized constructions.

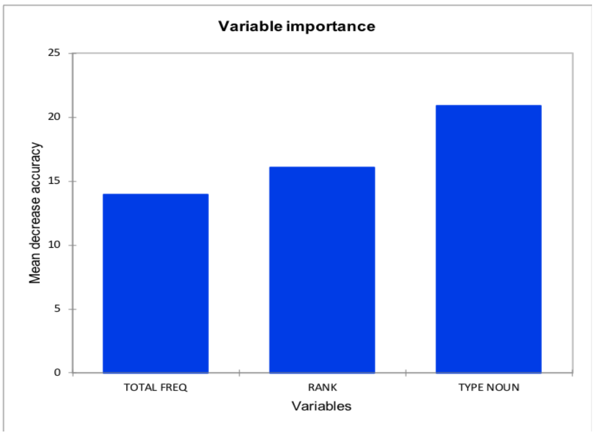


Figure 4.5: Mix variable importance assessment for true light/vague action verb prediction: corpora-based `Total_Frequency` and LVCs `Rank` (quantitative) and `Type_Noun` elements (qualitative).

Accordingly, at least there are three linguistic implications. Initially, noun semantics as a key differentiator. The prominence of `TYPE_NOUN` highlights the significance of noun semantics in distinguishing between true light verbs and vague action verbs. This suggests that the semantic class of the noun plays a crucial role in determining the choice of verb and the overall interpretation of the LVC. Besides, concerning frequency and conventionality. The moderate importance of `RANK` and the mixed impact of `TOTAL_FREQ` suggest that the frequency and conventionality of LVCs contribute to verb classification, but their influence varies depending on the verb type. True light verbs might be more sensitive to frequency and ranking, while vague action verbs might exhibit greater flexibility in their usage patterns. Lastly, interplay of quantitative and qualitative factors. The analysis underscores the complex interplay between quantitative factors (frequency and ranking) and qualitative factors (noun type) in shaping the behavior and interpretation of LVCs. This highlights the need for a multifaceted approach to understanding these constructions, considering both the statistical properties and the semantic nuances of their components.

4.4.2 Type of nouns as the predictor of verb element

According to the evaluation of type of noun, i.e. stative and eventive nouns, the several key observations can be revealed, and linguistic implications can be formulated. Table 4.15 (Part A) presents the results of a variable importance analysis. This analysis focuses on evaluating the predictive power of two noun types—Psychological State (Ps) and Physiological State (Ph)—in distinguishing between true light verbs (TL) and vague action verbs (VA).

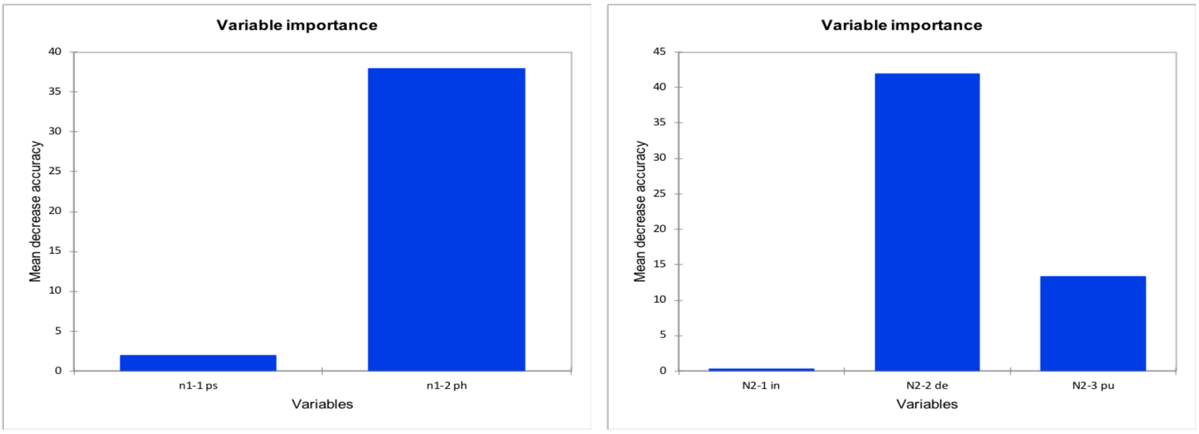
As the first key observations, physiological state (`Ph`) as a dominant predictor. The '`Ph`' (Physiological State) noun type emerges as the most influential predictor, with a strikingly high mean decrease in accuracy (37.863). Its impact is substantial for both '`Tl`' (27.164) and '`Va`' (29.051) classes, highlighting its critical role in differentiating between true light verbs and vague action verbs within Indonesian LVCs, as illustrated in Figure 4.6(a). This suggests that the presence of a physiological state noun is a strong indicator of the verb type within the LVC. Second, negative impact of psychological state (`Ps`). Interestingly, the permutation of the '`Ps`' (Psychological State) noun type results in a negative mean decrease in accuracy (1.916). This implies that disrupting this noun type actually improves the model's ability to classify verbs, particularly for true light verbs (tl: -4.056). This counter-intuitive result might suggest that psychological state nouns are less reliable predictors of verb type, or they might be more evenly distributed across both TL and VA verbs, making their disruption less detrimental to the model's performance.

Therefore, there are at least two linguistic implications. Regarding the physiological states and verb classification, the prominence of '`Ph`' nouns in predicting both TL and VA verbs underscores their significance in shaping the semantic and syntactic structure of LVCs. The strong association with both verb types suggests that nouns denoting physiological states play a crucial role in influencing the choice of verb and the overall interpretation of the construction. Also, about complex role of psychological states, the negative impact of '`Ps`' noun permutation suggests a more nuanced relationship between these nouns and the TL/VA distinction. Further analysis is needed to understand why disrupting this noun type improves the model's performance. It's possible that psychological state nouns exhibit greater flexibility in combining with different verb types, or that other linguistic features are more salient in predicting verb classification in LVCs containing these nouns.

Table 4.15: `TYPE_NOUN` for predicting the status of true light or vague action verbs.

PART A				PART B			
Variables	**tl**	**va**	**Mean decrease accuracy**	**Variables**	**tl**	**va**	**Mean decrease accuracy**
n1-1 Ps	-4.056	4.379	1.916	N2-1 In	7.228	-6.015	0.248
n1-2 Ph	27.164	29.051	**37.863**	N2-2 De	49.119	25.886	**41.914**
				N2-3 Pu	27.976	-9.379	13.289

Furthermore, Table 4.15 (Part B) presents the results of a variable importance analysis derived from a Random Forest model. The central focus here is to assess the predictive power of different types of eventive nouns, categorized as Indefinite Process (In), Definite Process (De), and Punctual (Pu), in distinguishing between true light verbs (TL) and vague action verbs (VA). As the first notable observations, Definite Process Nouns (De) as the Dominant Predictor, as illustrated in Figure 4.6(b). The '`De`' (Definite Process) noun type of eventive process emerges as the most influential predictor, with a strikingly high mean decrease in accuracy (41.914). It significantly impacts both 'tl' (49.119) and 'va' (25.886) classes, underscoring its crucial role in differentiating between true light verbs and vague action verbs within Indonesian LVCs. This suggests that the presence of a definite process noun, representing a specific and identifiable action or event, strongly influences the choice of verb type within the LVC.



(a) (b)

Figure 4.6: Evaluation results on `TYPE_NOUN` for predicting the status of true light or vague action verbs.

Second, punctual nouns (`Pu`) with Moderate Importance. '`Pu`' (Punctual) nouns, signifying momentary events or actions, demonstrate moderate predictive power, with a mean decrease in accuracy of 13.289. While their impact on predicting true light verbs (tl: 27.976) is substantial, their influence on predicting vague action verbs is negative (-9.379). This negative value suggests that permuting punctual nouns actually improves the model's ability to identify vague action verbs. This counter-intuitive result might indicate a complex relationship between punctual nouns and the VA category, potentially revealing a tendency for these nouns to co-occur less frequently with vague action verbs. Third, Indefinite Process Nouns (In) with Minimal Impact. 'In' (Indefinite Process) nouns, representing ongoing or habitual actions, exhibit the lowest overall importance (0.248). Their impact on both 'tl' (7.228) and 'va' (-6.015) is relatively small, with a negative value for 'va' suggesting a slight improvement in prediction accuracy when these nouns are permuted. This indicates that indefinite process nouns might be less reliable predictors of verb type within LVCs.

Accordingly. the implications of linguistic aspect from the observations are these three items. As of definite processes and verb choice, the prominence of 'De' nouns in predicting both TL and VA verbs highlights their significance in shaping the semantic and syntactic structure of LVCs. The strong association with both verb types suggests that definite process nouns play a crucial role in influencing the choice of verb and the overall interpretation of the construction. Also, reaction of Punctual Nouns and Vague Action Verbs. The negative impact of 'Pu' noun permutation on predicting vague action verbs suggests a potential aversion between these noun types. This could indicate that punctual nouns, representing momentary events, might be less compatible with the broader semantic scope often associated with vague action verbs. Lastly, Indefinite Processes and Flexibility. The minimal impact of 'In' nouns suggests that they might exhibit greater flexibility in combining with both TL and VA verbs. This could reflect the less specific and more durative nature of indefinite processes, allowing for a wider range of verb choices within LVCs.

4.4.3 Specific classes within stative and eventive noun as the predictor for verb-element selection

The subsequent question that arises in our discussion of the verbal element is whether the specific classes within stative and eventive nouns exert an influence on verb predictions. This inquiry is of paramount importance because, as qualitative variables, these classes often carry distinct semantic and syntactic features that could potentially impact the prediction of the verb type within Indonesian LVCs. The underlying assumption is that the inherent semantic properties of a noun, whether it denotes a state or an event, might predispose it towards certain types of verbs. Furthermore, the syntactic behavior of different noun classes could also play a role in determining the compatible verb choices. By examining the relationship between noun classes and verb selection, we can gain treasured comprehensions into the involved mechanisms governing the formation and interpretation of LVCs in Indonesian.

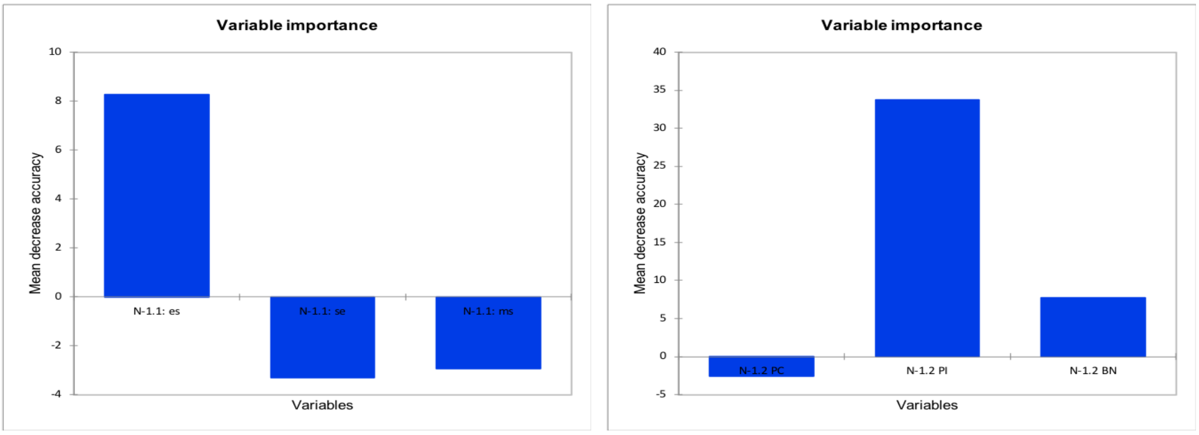
Based on the evaluation test, Table 4.16 (Part A) presents the results of a variable importance analysis. The focus here is to assess the predictive power of psychological state nouns, categorized into Emotional states (Es), Sensations (Se), and Mental states (Ms), in distinguishing between true light verbs (TL) and vague action verbs (VA). Based on this evaluation, several observations are as follows. First, Emotional States (Es) as a Strong Predictor. The 'Es' (Emotional states) sub-domain emerges as a significant predictor, with a notably high mean decrease in accuracy (8.270). Its impact is relatively balanced across both 'tl' (7.499) and 'va' (8.890), suggesting that the presence of a psychological state noun denoting an emotion is informative in distinguishing between both true light and vague action verbs. Second, negative impact of Sensations (Se) and Mental States (Ms). Surprisingly, the permutation of these sub-domains leads to a negative mean decrease in accuracy, as illustrated in Figure 4.7(a). This implies that disrupting these noun types actually improves the model's ability to classify verbs. This counter-intuitive result might indicate that these noun categories are less reliable predictors of verb type or might be more evenly distributed across both TL and VA verbs, making their disruption less detrimental to the model's performance.

Table 4.16: PSYCHOLOGICAL_STATE and PHYSIOLOGICAL_STATE for predicting the status of true light or vague action verbs.

PART A				PART B			
Variables	**tl**	**va**	**Mean decrease accuracy**	**Variables**	**tl**	**va**	**Mean decrease accuracy**
N-1.1: Es	7.499	8.890	**8.270**	N-1.2 Pc	-2.559	-2.251	-2.530
N-1.1: Se	-3.317	-3.143	-3.297	N-1.2 Pi	30.671	35.737	**33.684**
N-1.1: Ms	-3.040	-2.576	-2.926	N-1.2 Bn	1.200	11.971	7.706

Accordingly, the linguistic implications are in two folds. As of emotional states and verb classification, the strong association of 'Es' nouns with both TL and VA verbs suggests that emotional states are expressed through a diverse range of verbal constructions in Indonesian LVCs. The significant impact on both verb types indicates that the presence of an emotional state noun provides valuable information for distinguishing between them, possibly due to the nuanced semantic interplay between verbs and emotions. Additionally, there is a complex relationship with other sub-domains. The negative impact of 'Se' and 'Ms' permutations on the model’s accuracy suggests a more nuanced relationship between these noun categories and the TL/VA distinction. This could indicate that these noun types are less reliable predictors of verb type, or that they might be more flexibly combined with both TL and VA verbs, making their disruption less detrimental to the model’s performance.

Furthermore, based on the predictor evaluation of second class of stative noun, Table 4.16 (Part B) presents the results of a variable importance analysis. The central focus here is to evaluate the predictive power of physiological state nouns, categorized into Physical sensations (Pc), Physical states (Pi), and Basic physiological needs (Bn), in distinguishing between true light verbs (TL) and vague action verbs (VA). As the first observations, physical states (Pi) as the dominant predictor. The ‘Pi’ (Physical states) sub-domain emerges as the most influential predictor, with a strikingly high mean decrease in accuracy (33.684). Its substantial impact on both ‘tl’ (30.671) and ‘va’ (35.737) classes underscores its crucial role in differentiating between true light verbs and vague action verbs within Indonesian LVCs. This suggests that the presence of a physiological state noun plays a pivotal role in determining the verb type within the LVC.



(a) (b)

Figure 4.7: Evaluation results on PSYCHOLOGICAL_STATE and PHYSIOLOGICAL_STATE for predicting the status of true light or vague action verbs.

Second, basic physiological needs (Bn) with moderate importance. ‘Bn’ (Basic physiological needs) demonstrates a moderate level of predictive power, with a mean decrease in accuracy of 7.706. While its impact on predicting true light verbs (tl: 1.200) is relatively low, its influence on predicting vague action verbs (va: 11.971) is considerably higher. This indicates that nouns denoting basic physiological needs are particularly informative in identifying vague action verbs within LVCs. Third, negative impact of physical sensations (Pc), surprisingly, the permutation of the 'Pc' (Physical sensations) sub-domain results in a negative mean decrease in accuracy (-2.530). This implies that disrupting this noun type actually improves the model's ability to classify verbs. This counter-intuitive result might suggest that physical sensation nouns are less reliable predictors of verb type or might be more evenly distributed across both TL and VA verbs, making their disruption less detrimental to the model's performance.

Accordingly, the implications are as follows. Regarding the physical states as key differentiators, as illustrated in Figure 4.7(b), the prominence of 'Pi' nouns in predicting both TL and VA verbs highlights their significance in shaping the semantic and syntactic structure of LVCs. The strong association with both verb types suggests that nouns denoting physical states play a crucial role in influencing the choice of verb and the overall interpretation of the construction. Additionally, about the basic needs and vague action verbs, the notable impact of 'Bn' nouns, particularly on predicting vague action verbs, indicates a potential affinity between these nouns and verbs that carry more semantic weight. This could reflect the tendency for LVCs expressing basic physiological needs to involve actions or processes that are less easily captured by true light verbs with their more grammatical function. Lastly, Complex Role of Physical Sensations. The negative impact of 'Pc' noun permutation suggests a complex relationship between these nouns and the TL/VA distinction. Further analysis is needed to understand why disrupting this noun type improves the model's performance.

In addition to the class of STATIVE_NOUN, Table 4.17 (Part A) presents the results of a variable importance analysis of the class of EVENTIVE_NOUN derived from the same machine learning model. The central focus here is to evaluate the predictive command of DEFINITE_PROCESS nouns, specifically categorized into Physical Activities (Pa), Verbal Communication (Vc), and Cognitive Process (Co), in distinguishing between true light verbs (TL) and vague action verbs (VA). According to the results of calculation, three key observations are as follows. First, relatively balanced importance. The mean decrease in accuracy values for all three noun sub-domains are notably close, ranging from 9.543 to 12.241. This suggests that all three types of definite process nouns—physical activities, verbal communication, and cognitive processes—play a relatively balanced and significant role in distinguishing between true light verbs and vague action verbs within Indonesian LVCs.

Second, Physical Activities (Pa) as a slight frontrunner. While the differences are not substantial, 'Pa' exhibits the highest mean decrease in accuracy (12.241), indicating its marginally stronger predictive power compared to the other two sub-domains. This implies that the presence of a definite process noun denoting a physical activity might offer a slightly more reliable cue for differentiating between TL and VA verbs. Third, Consistent Impact Across Verb Types. For each noun sub-domain, the "mean decrease in accuracy" values for 'TL' and 'VA' are relatively close. This suggests that the predictive power of each noun type is fairly consistent across both true light verbs and vague action verbs.

Table 4.17: DEFINITE_PROCESS (part a) and PUNCTUAL_NOUNS (part b) for predicting the status of true light or vague action verbs.

PART A

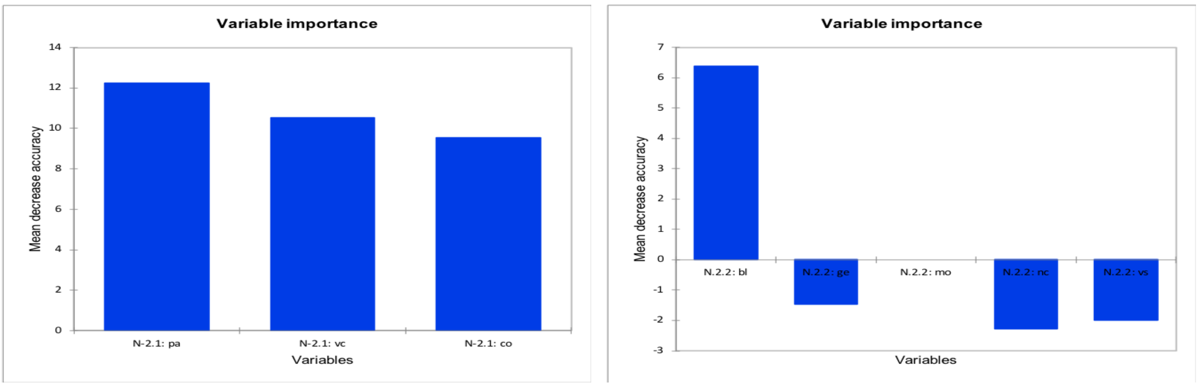
Variables	tl	va	Mean decrease accuracy
N-2.1: pa	12.335	11.634	12.241
N-2.1: vc	11.424	7.523	10.532
N-2.1: co	11.223	3.280	9.543

PART B

Variables	tl	va	Mean decrease accuracy
N.2.2: bl	5.351	0.718	6.389
N.2.2: ge	-2.370	-0.605	-1.453
N.2.2: mo	0.000	0.000	0.000
N.2.2: nc	-1.021	-1.112	-2.282
N.2.2: vs	1.955	-3.765	-1.991

For completing the observations, the linguistic implications are as follows. In respect to semantic and syntactic interplay, the relatively balanced importance of the three noun sub-domains suggests that the distinction between true light verbs and vague action verbs is not solely determined by the semantic category of the noun. The interplay between the semantics of the noun and the syntactic and semantic properties of the verb contributes to the overall interpretation and classification of the LVC. Accordingly, as of physical activities as a potential differentiator, the slightly higher importance of 'Pa' hints at a potential tendency for physical activity nouns to co-occur more frequently or distinctively with either true light verbs or vague action verbs. Further analysis of the specific verbs and constructions within this sub-domain could shed light on the linguistic factors that contribute to this pattern.

Moreover, regarding the PUNCTUAL_NOUNS for predicting the status of true light or vague action verbs within Indonesian LVCs, Table 4.17 (Part B) displays the results of a variable importance analysis, stemming from a Random Forest model. The primary goal here is to assess the predictive power of punctual process nouns, categorized into Blow (Bl), Gesture (Ge), Motion (Mo), Nonverbal communication (Nc), and Vocal sounds (Vs), in distinguishing between true light verbs (TL) and vague action verbs (VA) within Indonesian LVCs. Based on the data, several notes are as follows. First, Blow (Bl) as a strong predictor. The 'Bl' (Blow) sub-domain emerges as the most influential predictor, with the highest mean decrease in accuracy (6.389). Its substantial impact on 'Tl' (5.351) suggests that the presence of a punctual noun denoting a blow is a strong indicator of a true light verb in the LVC. The relatively lower impact on ‘Va’ (0.718) further supports this notion. Second, Negative Impact of Gesture (Ge), Nonverbal communication (Nc), and Vocal sounds (Vs). Surprisingly, the permutation of these sub-domains leads to a negative mean decrease in accuracy. This implies that disrupting these noun types actually improves the model’s ability to classify verbs, suggesting a complex or potentially inverse relationship between these noun categories and the TL/VA distinction. Third, Motion (Mo) with No Impact. The zero values for both ‘Tl’ and ‘Va’ in the ‘Mo’ (Motion) sub-domain indicate that this noun type has no discernible influence on predicting verb types within LVCs.



(a) (b)

Figure 4.8: Evaluation results on DEFINITE_PROCESS and PUNCTUAL_NOUNS for predicting the status of true light or vague action verbs.

Notably, some linguistic implications can be deduced form this calculation. Concerning the nouns of blow and true light verbs, the strong association between ‘Bl’ nouns and true light verbs suggests that these nouns tend to co-occur with verbs that have a more grammatical function, allowing the noun to carry the primary semantic load in the LVC. This aligns with the nature of true light verbs, which often contribute less to the overall meaning of the construction. Additionally, as of complex relationship with other sub-domains, the negative impact of ‘Ge’, ‘Nc’, and ‘Vs’ permutations on the model's accuracy suggests a more nuanced relationship between these noun categories and the TL/VA distinction. This could indicate that these noun types are less reliable predictors of verb type or that they might be more evenly distributed across both TL and VA verbs, making their disruption less detrimental to the model's performance. Lastly, neutrality of the nouns of motion. The lack of predictive power for ‘Mo’ nouns suggests that they might not be strongly associated with either true light verbs or vague action verbs, potentially indicating their flexibility in combining with different verb types.

4.5 Resume

Chapter 4 presents an in-depth analysis of the verb element within Indonesian LVCs, grounded in aktionsart theory and expanded through both qualitative linguistic evaluation and quantitative modeling. The chapter introduces two principal verb types—True Light Verbs (TLVs) and Vague Action Verbs (VAVs)—classified based on their temporal specification, degree of semantic bleaching, and combinatorial behavior with noun elements. TLVs, examined in §4.1, exhibit aspectual underspecification and primarily function as grammatical scaffolds, showing a tendency toward [-durative], [-frequentative], and [-intensive] values. The chapter distinguishes between optimal and suboptimal TLVs, where optimal TLVs (e.g., *melakukan, memberikan*) exhibit high type-pairing rates and distributional versatility, while suboptimal TLVs are more lexically constrained. In §4.2, VAVs are examined as semantically lighter but more context-sensitive verbs. They show variable [+]/[-] aktionsart values and are further divided into optimal and suboptimal types based on co-occurrence range with nouns. Section 4.3 applies machine learning techniques—including Naïve Bayes and Random Forest models—to assess the predictive power of frequency data and grammatical parameters. Notably, NB classifiers outperform other models in predicting optimal type-pairings, while Random Forest regression reveals that semantic parameters (S-1 and S-2) and corpus frequency patterns (especially in ILCC and IDC) play pivotal roles in verb classification. The chapter concludes with a comprehensive discussion (§4.4), demonstrating that noun-types—particularly definite process and physiological state nouns—strongly predict the verb type. This approach affirms that verb behavior in LVCs is shaped by an interplay of aspectual neutrality, combinatorial selectivity, corpus distribution, and grammatical context.

Chapter 5

Noun element within Indonesian LVCs

5.0 Introduction

A pivotal component of analyzing Indonesian LVCs resides in the examination of their noun elements. The semantic properties of the noun heads within LVCs significantly influence the construction's overall meaning and usage patterns. The frequency and distribution analysis also reveal that there is a need for a detailed analysis of the noun element in those constructions. According to the current analysis of the aktionsart, Indonesian LVCs at least indicates the existence of two primary categories of noun element, as illustrated in Figure 5.1, namely *stative nouns* (§5.1) and *eventive nouns* (§5.2). The stative nouns can be further classified into *psychological states* and *physiological states*, while eventive nouns can be classified into *indefinite process nouns*, *definite process nouns*, and *punctual nouns*.

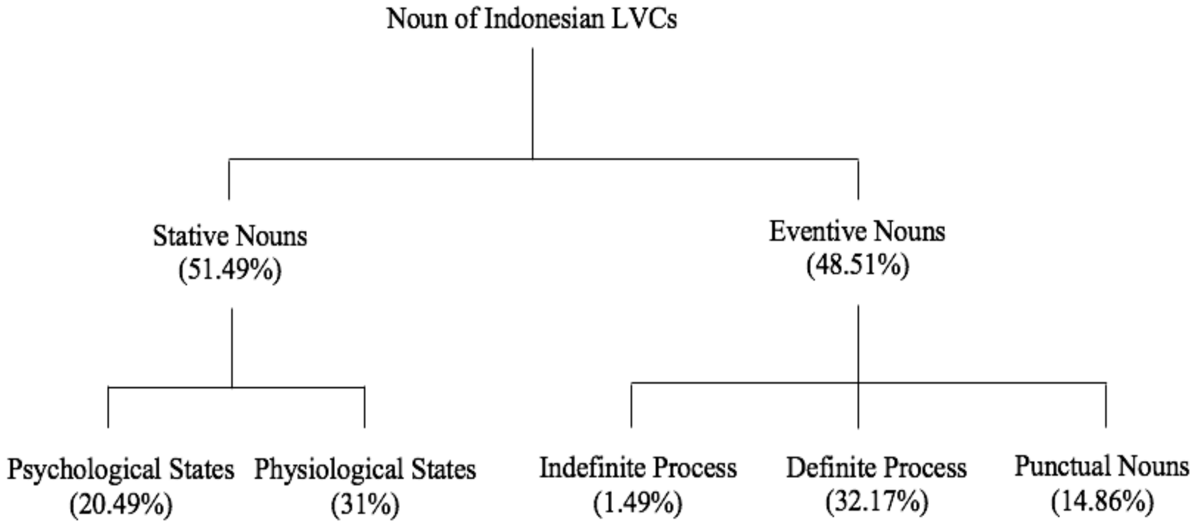

Noun of Indonesian LVCs
Stative Nouns (51.49%)
Eventive Nouns (48.51%)
Psychological States (20.49%)
Physiological States (31%)
Indefinite Process (1.49%)
Definite Process (32.17%)
Punctual Nouns (14.86%)

Figure 5.1: Hierarchical classification of noun elements in Indonesian LVC, with percentage division across categories.

In general, the KDE plot as in Figure 5.2 presents a visualization of the distribution of total-LVCs values, segmented by noun type (Eventive LVCs (Ev) and Stative LVCs (St)), across three distinct frequency clusters. This visualization employs kernel density functions to estimate the probability density function of the total-LVCs variable, thereby providing a smoothed representation of the underlying data distribution. The division into clusters, originated from a prior K-means clustering analysis, allows for a comparative assessment of how the distribution of total-LVCs varies as a function of both noun type and cluster membership. The use of separate subplots for each cluster facilitates a clear, side-by-side comparison, highlighting both commonalities and divergences in the distributional characteristics.

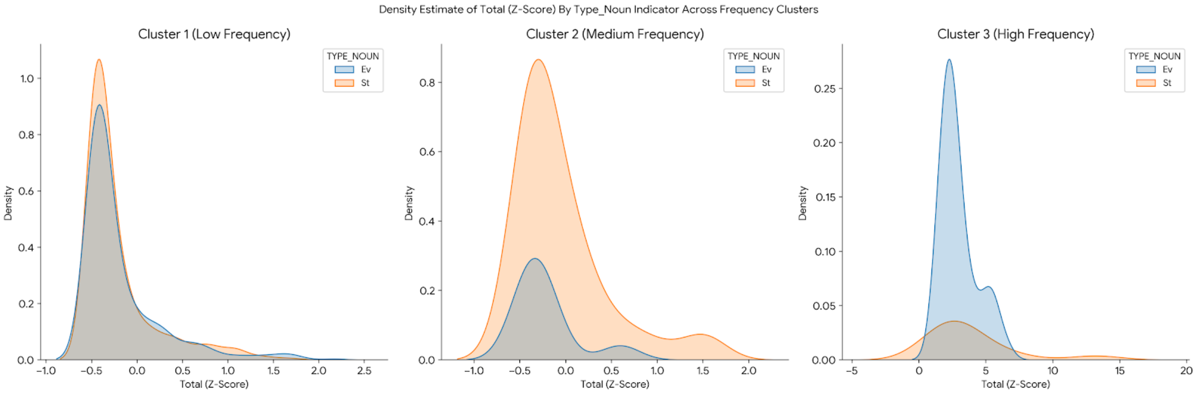


Figure 5.2: Distribution of stative and eventive LVCs across frequency's clusters.

Cluster 1, representing the Low Frequency category, exhibits a notable concentration of data points at the lower end of the total-LVCs (Z-score) value range. Both Eventive (orange) and Stative (blue) LVCs demonstrate a rightward skew, indicating a greater probability of observing lower total values, with a long tail extending towards higher values. However, the Eventive LVCs display a more pronounced and narrower peak at the lower end of the total scale, suggesting that low-frequency occurrences are more characteristic of this noun type within this cluster. The Stative LVCs, while also skewed, show a slightly less pronounced and broader peak, indicating a greater variability in total values, even within this low-frequency context. The separation in the density curves suggests a discernible difference in the distributional behavior of the two noun types even within the lowest frequency cluster.

In Cluster 2, classified as Medium Frequency, the distribution of total-LVCs values demonstrates a shift towards higher ranges compared to Cluster 1, suggesting a greater prevalence of mid-range total values. The Stative LVCs, however, display a more noticeable change, with a broader distribution and a less pronounced peak shifted toward slightly higher values. This suggests that as the frequency of occurrence increases, Stative LVCs tend to exhibit a wider range of total values. The differences between the two noun types are less stark in this cluster, indicating an increased overlap in their distributional characteristics at medium frequencies.

Cluster 3, representing the High Frequency, showcases a further shift in the distribution of total-LVCs values towards the higher end of the spectrum, particularly for Eventive LVCs. Both Eventive and Stative LVCs display a more symmetrical distribution compared to the previous clusters, indicating a more even spread of total-LVCs values at higher frequencies. However, Eventive LVCs exhibit a peak that is higher than the peak of Stative LVCs in this cluster, suggesting that certain total values are more common for Eventive LVCs even at high frequencies. The Stative LVCs show a more defined peak in this cluster compared to the broader distribution in lower frequency clusters, although their overall spread is still considerable.

In conclusion, the KDE plots reveal a trend of increasing total-LVCs values as we move from Cluster 1 (Low Frequency) to Cluster 3 (High Frequency), although the shift isn't strictly monotonic for both Eventive and Stative LVCs. While Eventive LVCs tend to exhibit a more pronounced and narrower peak in their distribution, particularly at lower frequencies (Cluster 1), this peak appears broaden and shift slightly towards higher values in Cluster 2 and becomes less distinct in Cluster 3. Stative LVCs demonstrate a broader distribution compared to Eventive LVCs within each cluster, indicating greater variability in their total values. The differences between the two noun types are most pronounced at lower frequencies (Cluster 1), suggesting that the stativity and eventivity of the noun element exerts a greater influence on the distribution of total values when the overall frequency of occurrence is low. As frequency increases (especially from Cluster 2 to Cluster 3), the central tendencies of the two noun types appear to converge, but Eventive LVCs still show a more concentrated distribution while Stative LVCs maintain a wider spread.

5.1 Stative nouns

This present analysis delves specifically into LVCs featuring stative nouns with a particular emphasis on those conveying psychological (§5.1.1) and physiological states (§5.1.2). Due to the potential for these categories to exhibit distinct combinatory patterns within the Indonesian language (*see* Figure 5.4), the present analysis has opted to examine them individually. This granular method allows for a more nuanced understanding of how these specific stative noun classes interact with light verbs to create meaning within LVCs. By isolating these categories, we can identify any potential constrains that govern their co-occurrence within these constructions.

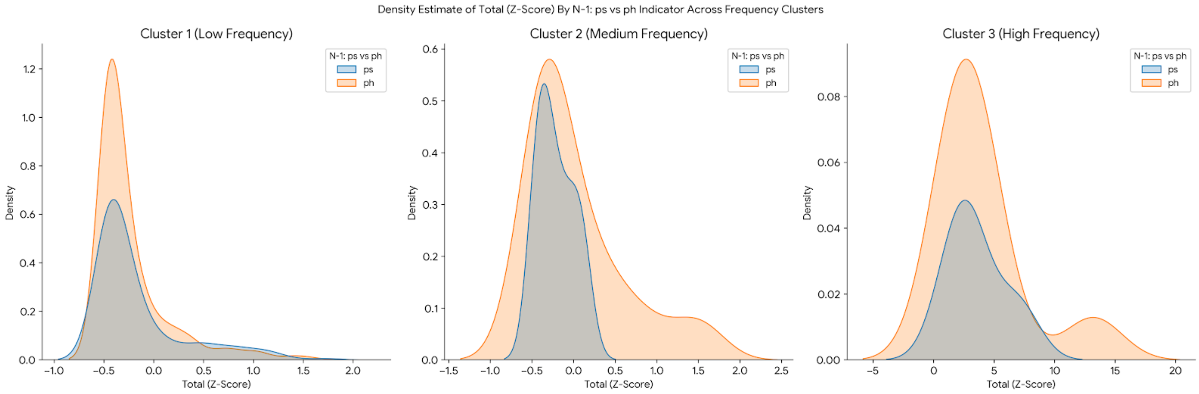


Figure 5.3: Density estimate of psychological and physiological noun within Indonesian LVCs.

In general, Figure 5.3 represents the density distribution of a total score across three distinct frequency clusters, stratified by indicators of psychological (Ps) and physiological (Ph) states. KDE is employed to provide a smoothed approximation of the underlying probability density function, offering insights into the likelihood of observing specific total score values within each cluster. The x-axis represents the total score, while the y-axis indicates the density, reflecting the relative frequency of occurrence. Each cluster is presented in separate subplot, facilitating a comparative analysis of the distributional characteristics across varying frequency bands. Notably, the differentiation between Ps and Ph indicators is achieved through distinct colored areas (orange for Ps and blue for Ph) within each cluster's KDE plot, allowing for a nuanced examination of how these states modulate the total score distribution.

In Cluster 1 (Low Frequency), the distribution for the psychological (Ps) indicator (orange) appears unimodal with a peak centered around a lower total score value (approximately 0). There is a slight shoulder or skewness towards higher values, but not a clearly distinct secondary mode. The physiological (Ph) indicator (blue) in Cluster 1 also appears largely unimodal, with its peak situated at a notably higher total score range (around 1-2) compared to the Ps indicator. While there is some overlap between the tails of the two distributions, the primary densities are clearly separated. Indicating that in this low-frequency context, psychological and physiological nouns tend to be associated with different ranges of total scores.

In the medium-frequency cluster (Cluster 2), the psychological (Ps) indicator still appears somewhat bimodal, although the separation between the two potential peaks is less pronounced compared to Cluster 1. While the primary peak of the ‘Ps’ distribution is centered around a moderate total score, suggesting a concentration of scores in this range, the presence of a smaller shoulder or secondary peak indicates some heterogeneity remains. The physiological (Ph) indicator in Cluster 2 maintains a broader distribution compared to ‘Ps’, with a less clearly defined peak and a noticeable skew towards higher total scores, suggesting greater variability in total scores associated with physiological states at this frequency level. The reduction in the prominence of the lower peak for ‘Ps’ compared to Cluster 1 suggests a potential attenuation of the distinct low-score pattern for psychological nouns at this intermediate frequency. The observation that the distributions in Cluster 2 are different from both Cluster 1 and Cluster 3 implies that the influence of frequency on score distribution is not linear, and that intermediate frequency ranges may indeed attenuate or reshape the distinct score patterns seen at lower frequencies.

Cluster 3, representing the high-frequency cluster, demonstrates a degree of convergence in the distribution patterns of psychological (Ps) and physiological (Ph) indicators, particularly in their central tendencies. While the ‘Ps’ distribution appears bimodal with a primary peak around a moderately high total score and a secondary, smaller peak at a lower score, the ‘Ph’ distribution exhibits a more clearly unimodal shape with a peak at a higher total score range compared to the primary peak of ‘Ps’. Therefore, stating that both distributions exhibit a unimodal shape peaking at a relatively high total score range is not entirely accurate for ‘Ps’. This suggests that at high frequencies, while higher total scores are generally observed for both types, the relationship might be more complex for psychological indicators, potentially involving two distinct patterns of association. The increased overlap in the distributional characteristics compared to Clusters 1 and 2 implies that the differentiation between ‘Ps’ and ‘Ph’ becomes less pronounced in this high-frequency context in terms of their central tendencies, potentially indicating a ceiling effect or a heightened sensitivity to factors influencing total scores at elevated frequencies. However, the ‘Ps’ distribution still shows a wider spread than the primary peak of the ‘Ph’ distribution, indicating that the reduction in dispersion is more evident for physiological indicators.

In conclusion, the comparative analysis of the KDE plots reveals a complex interplay between frequency, psychological (Ps) and physiological (Ph) states, and total score distribution. Specifically, Cluster 1 shows evidence of bimodality, particularly in the distribution of psychological nouns, suggesting two distinct ranges of lower total scores within this category. While physiological nouns in Cluster 1 exhibit a more unimodal distribution, their score are generally lower than those observed in higher frequency clusters. Cluster 2 displays a shift towards unimodality for both 'Ps' and 'Ph' nouns, with both distributions showing a noticeable increase in total scores compared to Cluster 1. Finally, Cluster 3 predominantly exhibits unimodal distributions for both 'Ps' and 'Ph' nouns, with majority of scores concentrated at the higher end of the total score range. This transition from potential bimodality at low frequencies to unimodality at higher frequencies suggests a dynamic relationship where lower frequency context might be associated with more differentiated total score patterns between psychological and physiological nouns, while higher frequency contexts demonstrate a convergence towards elevated total scores.

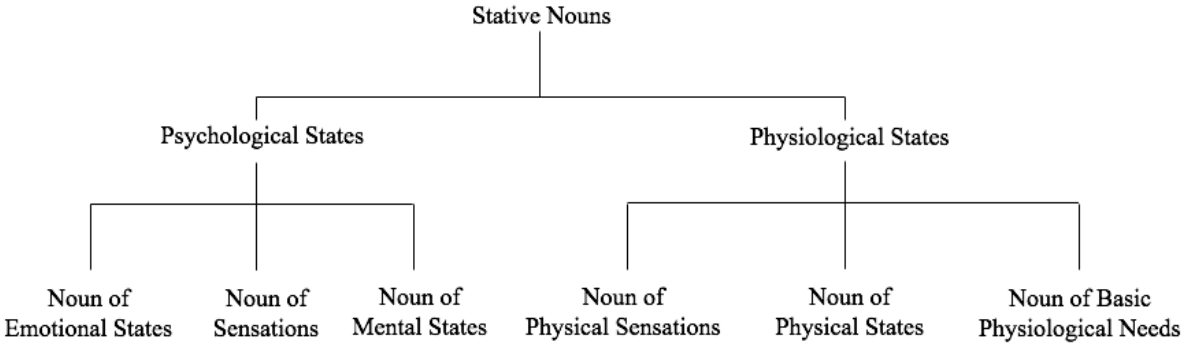


Figure 5.4: Classification of stative nouns within Indonesian LVCs.

5.1.1 Noun of psychological states

Focusing on psychological states within the category of stative nouns employed in LVCs, I can further subdivide them into distinct subcategories. As illustrated in Table 5.4, the data reveals a tapestry of psychological states expressed through these constructions. Nouns encompassing emotional states, such as '*keberanian'* (courage) and '*kehormatan'* (honor), readily combine with light verbs to depict emotional experiences. Similarly, sensations, as exemplified by '*rasa*' (feel) and '*kepuasan'* (satisfaction), readily participate in LVCs. Finally, the realm of mental states is also well-represented, with nouns like '*kesadaran*' (awareness) and '*arti*' (meaning) forming part of these constructions. This breakdown of psychological state subcategories within stative nouns employed in LVCs allows for a more thorough analysis of their semantic properties.

Table 5.1: Notable list of the Indonesian LVCs containing psychological state nouns.

No.	LVC	PMW	No.	LVC	PMW
1.	*memiliki daya tarik* 'has attraction'	4912.627	11.	*memberikan kepuasan* 'give a satisfaction'	1466.621
2.	*berhutang budi* 'to be indebted to someone'	659.075	12.	*membawa kebaikan* 'put fine'	665.956
3.	*mengangkat sumpah* 'to take an oath'	285.037	13.	*mendapat cela* 'to give embarrassment'	4.671
4.	*membersihkan nama* 'to clear one's name'	222.354	14.	*membuat janji* 'make a promise'	593.077
5.	*menjaga kehormatan* 'defend honor'	1030.954	15.	*memiliki iman* 'to have faith in'	349.628
6.	*memiliki keberanian* 'give courage'	738.885	16.	*mati rasa* 'to feel nothing'	1139.845
7.	*menindak tegas* 'to severely crack down'	3566.199	17.	*memberikan arti* 'to give meaning'	331.848
8.	*memberikan pengaruh* 'to come into effect'	3499.297	18.	*tersentuh hatinya* 'to be touched'	101.257
9.	*mengambil sikap* 'takes a stand'	2943.437	19.	*memberi restu* *'give permission'*	152.237
10.	*menjadi terkenal* 'become famous'	740.392	20.	*menanamkan kesadaran* 'become aware'	274.087

Specifically, Table 5.1 presents a curated collection of Indonesian LVCs that incorporate nouns denoting psychological states. These LVCs offer a glimpse into how Indonesian lexicalizes complex psychological phenomena through the combination of verbs and nouns. The table further categorizes these nouns into three sub-domains: emotional states (Es), sensations (Se), and mental states (Ms), providing a framework for a more nuanced linguistic analysis.

5.1.1.1 Noun of emotional states

First subdivision of the psychological states within Indonesian LVCs is noun of emotional states or feelings. Based on the data of Table 5.1 (No 1-7), the first seven LVCs showcase a range of emotional states expressed through nominalization. The light verbs '*memiliki'* (to have) and '*memberikan'* (to give) are frequently employed to attribute emotional qualities or actions to an entity. For instance, '*memiliki daya tarik'* (to have attraction) and '*memiliki keberanian'* (to have courage) depict inherent qualities, while '*memberikan pengaruh'* (to come into effect) expresses the causative impact on emotions. The LVC '*berhutang budi*' (to be indebted to someone) exemplifies the cultural embedding of emotions, highlighting the complex interplay of gratitude and obligation, as well as '*menjaga kehormatan'* 'defend honor' and '*mengangkat sumpah'* (to take an oath).

(5.1) *Pikiran sadarnya menginginkan agar dia* ***memiliki keberanian*** *tersebut.*
mind his conscious desire that he possess courage that
'His conscious thoughts harbored a desire for him to have that courage.'

(5.2) *Sharon* ***mengangkat sumpah*** *sebagai perdana menteri.*
Sharon lifted an oath as prime minister
'Sharon took the oath of office as prime minister.'

(5.3) *Saya merasa* ***berhutang budi*** *kepada banyak orang.*
I feel indebted help to many people
'I feel a sense of gratitude/indebtedness towards many people.'

(5.4) *Dia selalu* ***menjaga kehormatan*** *dan nama baik suaminya.*
She always guards the honor and good name her husband
'She always upholds the honor and reputation of her husband.'

The four sample LVCs as in (5.1) to (5.4), namely '*memiliki keberanian*', '*mengangkat sumpah*', '*berhutang budi*', and '*menjaga kehormatan*', offer critical insight into the role of emotional and psychological nouns within Indonesian LVCs, particularly in relation to their stative properties. According to their inherent features, nouns such as '*keberanian*', '*budi',* and '*kehormatan*', are best characterized as [+durtivity] and [-dinamicity], representing internal states or ethical-psychological qualities rather than dynamic events. These nouns contribute sustained, atelic semantic content to the construction. In '*memikili keberanian*' (to have courage) and '*berhutang budi*' (to be indebted), the emotional state is durative and introspective, with little implication of change. Similarly, '*menjaga kehormatan*' (to guard/defend honor) frames a continuous moral effort anchored in enduring state. The one exception is '*mengangkat sumpah*' (to take an oath), which introduces a ritualistic event structure and may lean toward [+boundedness] and [+telicity] due its performative nature, suggesting a transitional point in status. Thus, while most emotional-state nouns in this set function as stative elements, contributing to aspectual stability, they also highlight how ritualized or socially embedded contexts can modulate noun stativity, especially when paired with event-defining verbs. In essence, these nouns function as the semantic core, carrying the inherent meaning of emotions like attractiveness, courage, or gratitude. Within the LVCs framework, they act as the semantic foundation upon which meaning is constructed. Light verbs, when coupled with these emotionally charged nouns, serve as grammaticalizing function, shaping the overall LVC but holding minimal semantic weight themselves.

5.1.1.2 Noun of sensations

Second subdivision of the psychological states within Indonesian LVCs is noun of sensations. According to Table 5.1 (No 8-14), the subsequent LVCs illustrate the linguistic representation of sensations, often intertwined with emotional experiences. LVC ‘*menindak tegas*’ (to severely crack down) and ‘*mengambil sikap*’ (to take a stand) depict actions driven by sensations or resulting in sensory experiences. The metaphorical use of ‘*mati rasa*’ (to feel nothing) illustrates the capacity of LVCs to encapsulate complex sensory and emotional states.

(5.5) *Salah satunya* ***menindak tegas*** *perjudian di masyarakat.*
one of them to act firm gambling in society
‘One of them is cracking down on gambling in society.’

(5.6) *Pihak terkait segera* ***mengambil sikap.***
parties related immediate to take action
‘The authorities should take immediate action.’

(5.7) *Kurasa sebagian tubuhku* ***mati rasa.***
I think part of my body numb feeling
‘I feel like part of my body has gone numb.’

As demonstrated in (5.5) to (5.7), these nouns embody the lexical essence of psychological sensations triggered by external stimuli. When incorporated into LVCs, they transform into the semantic cornerstone of the construction. Light verbs, bereft of their own significant meaning, serve primarily to grammaticalize the constructions. This symbiotic relationship allows the sensational noun to retain its core meaning and dictate the overall interpretation of the LVC. The spectrum of sensations encompassed by these nouns, ranging from psychological discomfort to auditory perception. Specifically, the three LVCs, namely ‘*mati rasa*’, ‘*mengambil sikap’*, and ‘*menindak tegas*’, exemplify realizations of stative nominal elements in LVCs. The construction ‘*mati rasa*’ (to go numb) contains the noun *rasa,* a prototypical sensation noun, which align with features [+durativity] and [-dynamicity], signifying an ongoing, internal state. The verb *mati* functions metaphorically to indicate the cessation of sensation but contributes no inherent dynamicity to the LVCs. In '*mengambil sikap*' (to take a stance), the noun *sikap* (attitude or stand) similarly denotes a stative psychological orientation, suggesting [+durative] and [-dynamicity] featues. Though the verb *mengambil* is agentive, its semantic contribution is minimal and supports predication rather than change. Thereafter, '*menindak tegas*' (to take firm action), while appearing more dynamic due to its collocational form, can involve institutional or procedural stativity depending on context—*tegas* (firmly) qualifies the manner rather than initiating an event. As such, the noun component (implied or reconstructed as *tindakan*) contribute a stative reading. Collectively, these demonstrate how sensation and attitudinal nouns structure the aspectual interpretation of the construction, grounding them in durative, non-telic states while the verb remains a syntactic facilitator.

5.1.1.3 Noun of mental states

Third subdivision of the psychological states within Indonesian LVCs is noun of mental states. Based on Table 5.1 (No 15-20), the final list LVCs delve into the realm of mental states, encompassing cognitive processes, beliefs, and intentions. LVC '*membuat janji*' (to make a promise) and '*memiliki iman*' (to have faith) showcase how LVCs capture commitments and belief systems. Constructions like '*menanamkan kesadaran*' (to become aware) and '*memberikan arti*' (to give meaning) illustrate the dynamic nature of mental states and their susceptibility to external influences. Delving into the realm of stative nouns within LVCs, we encounter a category encompassing both mental states. Mental state nouns, such as '*kesadaran*' (to become aware) or '*arti*' (meaning), capture the internal workings of the mind. Nouns of aptitude, exemplified by '*janji*' (promise), highlight inherent capabilities or dispositions. When integrated into LVCs, these nouns act as the semantic cornerstone. Light verbs, devoid their own independent meaning in this context, serve primarily to grammaticalize the constructions. Despite this grammatical function, the light verb does not diminish the inherent meaning of the mental state or aptitude noun. Instead, the LVC construction acts as a framework, allowing these core nouns to take stage and shape the overall meaning expressed.

(5.8) *Ia* ***membuat*** ***janji*** *dengan* *Timo.*
he/she makes promise with Timo
'He/she made a promise to Timo.'

(5.9) *Tuhan* *menghendaki* *agar* *kita* ***memiliki*** ***iman***.
God wills so that we posesses faith
'God desires us to have faith.'

(5.10) *Butuh* *waktu* *untuk* ***menanamkan*** ***kesadaran*** *kepada* *masyarakat.*
need time for instill awareness to society
'It takes time to raise awareness in society.'

(5.11) *Pembela* *jelas* ***memberikan*** ***arti*** *lain.*
defender clearly gives meaning other
'The defender clearly gives it a different meaning.'

Particularly, nouns of mental state constitute a subcategory of psychological states, characterized by internalized, cognitively grounded meanings. As exemplified in (5.8) to (5.11), LVCs '*membuat janji*', '*memiliki iman*', '*menanamkan kesadaran*', and '*memberikan arti*' exemplifies this type. The nouns *janji* (promise), *iman* (faith), *kesadaran* (awareness), and *arti* (meaning) convey [+durativity] and [-dynamicity], denoting mental state or belief systems that persist over time without requiring overt action. These LVCs are not constructed to narrate events, but rather than to express cognitive positioning or intentional states. The verbs (e.g., *membuat, memiliki*) contribute little aspectual force and instead enable grammatical realization. For instance, *menanamkan* introduces a metaphor of implantation but functions to reinforce the internalization of *kesadaran.* Overall, these nouns serve as semantic anchors, allowing the LVC to maintain a stable, atelic interpretation. This subclass within stative nouns highlights the importance of mental-cognitive grounding in shaping the temporal structure of Indonesian LVCs.

5.1.2 Noun of physiological states

Within the category of stative nouns employed in LVCs, differentiating between experiential meaning and physiological states is crucial. Experiential meanings encompass physical states and physical sensations. Physical states, exemplified by nouns like '*tangan*' (hand) or '*kecupan*' (a kiss), depict a material state and temporary condition within the experiencer. Conversely, physical sensations, such as '*kepanasan*' (hot) or '*aroma*' (a smell perceived), describe a more immediate and localized bodily perception. Separating these from physiological states, which refer to basic biological needs, is also vital. Noun like '*energi*' (energy) or '*nyawa*' (life) represent basic physiological needs, often with a more long-lasting impact on the experiencer's well-being. This distinction ensures a clearer understanding of how stative nouns, categorized by their experiential or physiological nature, contribute to the semantic richness of LVCs in Indonesian.

Table 5.2: Notable list of the Indonesian LVCs containing physiological state nouns.

No.	LVC	PMW	No.	LVC	PMW
1.	*menghela nafas* 'give a sigh'	2450.562	11.	*memasang taruhan* 'make a bet'	333.405
2.	*mencium bau* 'do smell	1984.910	12.	*melayangkan surat* 'to write a letter'	3319.987
3.	*mengeluarkan keringat* 'to sweat'	544.809	13.	*mengambil catatan* 'take notes'	16.073
4.	*membawa beban* 'carry a load'	380.618	14.	*campur tangan* 'to interfere'	9253.723
5.	*mengambil jarak* 'take distance'	324.113	15.	*mengalami demam* 'have a fever'	887.255
6.	*merasa kepanasan* 'be hot'	148.119	16.	*merasa pusing* 'become dizzy'	498.701
7.	*menebarkan aroma* 'have a smell perceived'	146.863	17.	*membutuhkan energi* 'require energy'	480.821
8.	*mendaratkan kecupan* 'give kiss'	53.240	18.	*meneteskan air mata* 'do cry'	247.819
9.	*membuka jalan* 'pave the way'	1852.261	19.	*menyambung hidup* 'to make a living'	508.546
10.	*menggelar dagangan* 'launches on the market'	190.460	20.	*mencabut nyawa* 'to take someone's life'	569.772

In particular, Figure 5.4 presents a breakdown of the sub-classification observed within physiological state nouns incorporated into Indonesian LVCs. These nouns are categorized into three distinct types: 'Physical sensations', 'Physical states', and 'Basic physiological needs'. Moreover, Table 5.2 presents a curated selection of Indonesian LVCs that incorporate nouns signifying physiological states. This collection serves as a valuable resource for investigating the linguistic representation and expression of human physical sensations, states, and basic needs within the Indonesian language. The table further categorizes these LVCs into three sub-domains, accompanied by frequency metrics from four examined corpora (ILCC, SLIC, IDC, and IWC), allowing for a subsequent analysis of their usage patterns.

5.1.2.1 Noun of physical sensation

First subdivision of the physiological states within Indonesian LVCs is nouns of physical sensation (Pc) (No 1-7 of Table 5.2). This category encompasses LVCs that express the subjective experience of bodily sensations. The predominant verb is '*merasakan*' (to feel), highlighting the perceptual nature of these experiences. Although seemingly simple, '*menghela nafas*' (give a sigh) captures a nuanced physiological response often associated with relief, exhaustion, or contemplation. Its high frequency across all corpora (Total Frequency: 48,790) underscores its ubiquity in Indonesian discourse. Next, '*mencium bau*' (to smell). This LVC denotes the sensory perception of odor. Its considerable frequency (39,519) suggests its importance in describing the environment and experiences. Also, '*mengeluarkan keringat*' (to sweat), this LVC signifies the physiological process of perspiration. Its lower frequency (10,847) compared to the previous two might indicate its more context-specific usage, perhaps related to physical exertion or emotional states. Lastly, '*membawa beban*' (carry a load), '*mengambil jarak*' (take distance), '*merasa kepanasan*' (be hot), '*menebarkan aroma*' (have a smell perceived), these LVCs, while less frequent, depict various physical sensations and actions, further enriching the linguistic repertoire for expressing bodily experiences.

The aforementioned nouns capture the lived experience of the body. Nouns of physical sensation focus on the immediate and localized perception within the body. When incorporated into LVCs, these semantically rich nouns co-occur with light verbs. The light verbs, however, assume a primarily grammatical role, shaping overall construction without eclipsing the core meaning carried by the noun of physical sensation.

(5.12) *Philip* ***merasa*** ***kepanasan*** *di* *tengah* *jerami.*
Philip feels hot in middle straw
'Philip felt very hot in the middle of the haystack.'

(5.13) *Saya* *masih* *dapat* ***mencium*** ***bau*** *busuk.*
I still can smell odor rotten
'I can still smell the foul odor.'

(5.14) *Jangan* ***membawa*** ***beban*** *terlalu* *berat* *di* *mobil.*
Don't carry burden too heavy in car
'Don't carry too heavy a load in the car.'

(5.15) *Ridwan* ***menghela*** ***nafas*** *dalam-dalam* *dan* *menghembuskannya.*
Ridwan inhaled breath very deep and exaheled it
'Ridwan took a deep breath and exhaled.'

(5.16) *Ternyata* *saya* *tidak* *perlu* ***mengeluarkan*** ***keringat*** *terlalu* *banyak.*
it turned out I not need to expel sweat too much
'It turned out I didn't need to sweat too much.'

(5.17) *Untuk itu* *kita* ***mengambil*** ***jarak***.
for that we take distance
'For that reason, we are keeping our distance.'

(5.18) *Jahe* *bisa* *membuat* *badan* ***menebarkan*** ***aroma*** *sedap.*
ginger can make body spread aroma pleasant
'Ginger can make someone's body give off a pleasant smell.'

In particular, as illustrated in (5.12) to (5.18), the noun elements in LVCs as '*merasa kepanasan*', '*mencium bau*', '*menghela nafas*', '*mengeluarkan keringat*', and '*menebarkan aroma*' exemplify the behavior of physiological stative nouns grounded in physical sensation. These nouns—*kepanasan* (overheated), *bau* (odor), *nafas* (breath), *keringat* (sweat), and *aroma* (scent)—encode sensory-affected states rather than discrete events. Based on their inherently features, the generally manifest as [+durative] and [-dynamicity], indicating conditions that persist over time without inherent change. While *nafas* and *keringat* may suggest episodic iteration in biological contexts, the LVC framing emphasizes experiential continuity rather than actionality. In *'mengambil jarak*' and '*membawa beban*', the stative interpretation arises from the abstracted sense of resistant or pressure rather than movement. Verbs such as *menghela, mengeluarkan,* and *menebarkan* add grammatical structure but do not overwrite the noun's temporal contribution. These constructions illustrate how physiological sensations are lexically encoded as internal, durative states, reinforcing the noun's central role in shaping LVC aspectuality.

5.1.2.2 Noun of physical states

Second subdivision of the physiological states within Indonesian LVCs is nouns of physical states (Pi) (No 8-14 of Table 5.2). This category comprises LVCs that describe objective physical states or conditions, often implying a degree of duration or impact on the individual. For instance, '*membuka jalan*' (pave the way), '*menggelar dagangan*' (launches on the market), *memasang taruhan* (make a bet), these LVCs, though metaphorically rooted in physical actions, primarily represent states or situations. Their varying frequencies reflect their diverse domains of application. Additionally, '*mendaratkan kecupan*' (give a kiss), while seemingly an action, this LVC also implies a physical state—the contact and pressure of lips on another surface. Its low frequency (1,060) suggests its specificity to intimate or affectionate contexts. Also, '*melayangkan surat*' (to write a letter), '*mengambil catatan*' (take notes), these LVCs, while involving physical actions, emphasize the resulting state—the existence of a letter or notes. Their frequencies suggest their relevance in communication and documentation. Lastly, '*campur tangan*' (to interfere), this high-frequency LVC (184,239) metaphorically extends a physical action to represent involvement or intervention in a situation.

(5.19) *Segera* *ia* ***membuka*** ***jalan*** *di* *depan.*
immediately he/she opened way in front
'Immediately he opened a way in front.'

(5.20) *Banyak* *orang* *berdagang* *dengan* ***menggelar*** ***dagangan*** *di* *pasar.*
many people trade by spreading merchandise in market
'Many people trade by spreading out their merchandise in the market.

(5.21) *Kuminta* *dia* *untuk* ***memasang*** ***taruhan*** *di* *angka 34* *atau* *43.*
I asked him to attach bet on number 34 or 43
'I asked him to place a bet on number 34 or 43.'

(5.22) *Lalu* *dia* ***mengelus*** ***pipi*** *Nania* *dan* ***mendaratkan*** ***kecupan*** *lembut.*
then he stroked cheek Nania and landed kiss soft
'Then he stroked Nania's cheek and landed a soft kiss.'

(5.23) *Kita* *akan* ***melayangkan*** ***surat*** *teguran.*
we will fly letter warning
'We will send a warning letter.'

(5.24) *Ia* ***mengambil*** ***catatan*** *itu* *dan* *membacanya*
he/she take notes that and tead it
'He took the notes and read them.'

(5.25) *Rakyat* *yang* ***turun*** ***tangan*** *membantu* *perjuangan.*
people who descend hand help struggle
'The people who took action helped the struggle.'

As exemplified in (5.19) to (5.25), the LVCs '*membuka jalan*', '*menggelar dagangan*', '*memasang taruhan*', '*mengelus pipi',* '*mendaratkan kecupan*', '*melayangkan surat*', '*mengambil catatan*', and '*turun tangan*' exemplify the use of physical-state nouns that denote tangible, manipulable entities or configurations. According to the feature's analysis, these nouns—*jalan, dagangan, taruhan, pipi, kecupan, surat, catatan,* and *tangan*—are best characterized by [+durative] and [-dynamicity], as they represent objects or bodily parts that persist as states or physical conditions. Their contribution to telicity and boundedness is variable, often shaped by discourse context or verb semantics. For example, *mengelus pipi* may suggest a brief physical action, but the noun *pipi* anchors the construction in bodily presence, not transition. Similarly, *turun tangan* metaphorically encodes intervention, yet *tangan* retains its stative core. In each case, the verb functions as a grammatical activator, while the noun serves as the aspectual nucleus, maintaining the LVC's overall durative and non-eventive interpretation. This pattern confirms the role of physical-stet nouns in grounding LVCs in spatial and embodied stability.

5.1.2.3 Noun of basic physiological needs

Third subdivision of the physiological states within Indonesian LVCs is nouns of basic physiological needs (Bn) (No 15-20 of Table 5.2). This category encompasses LVCs expressing fundamental human needs essential for survival and well-being. For instance, '*mengalami demam*' (have a fever), '*merasa pusing*' (become dizzy), these LVCs describe common physical ailments, indicating deviations from a state of well-being. Also, '*membutuhkan energi*' (require energy), '*meneteskan air mata*' (do cry), '*menyambung hidup*' (to make a living), *mencabut nyawa* (to take someone's life), these LVCs represent essential needs (energy, emotional release, sustenance, life itself) and their fulfillment or deprivation. In sum, it can be deduced that the verbs '*membutuhkan*' (to need) and '*mengalami*' (to experience) emphasize the necessity and impact of these physiological needs; the frequencies suggest the varying salience of these needs in different discourses; survival needs may be more prevalent in certain contexts than emotional or existential ones, and these LVCs contribute to the linguistic portrayal of human vulnerability and the fundamental requirements for existence.

(5.26) *Usaha untuk* ***menyambung hidup*** *harus tetap dilakukan.*
effort for connect life must still done
'The effort to make a living must still be made.'

(5.27) *Malaikat maut mencabut* ***nyawa seluruh*** *makhluk.*
angel death take life all beings
'The angel of death takes the life of all beings.'

(5.28) *Ibunya bercerita sambil* ***meneteskan airmata****.*
his/her mother tell story while dripping tears
'His/her mother told the story while shedding tears.'

(5.29) *Ingatlah bahwa tubuh anda* ***membutuhkan energi*** *setiap saat.*
remember that body your need energy every time
'Remember that your body need energy all the time.'

(5.30) *Saya* ***merasa pusing*** *pada saat pertama kali memakai kacamata ini.*
I feel dizzy at time first time wear glasses this
'I felt dizzy the first time I wore these glasses.'

(5.31) *Ini akan menyebabkan tubuh* ***mengalami demam.***
this will cause body experience fever.
'This will cause the body to experience a fever.'

In details, as demonstrated in (5.26) to (5.31), the six LVCs—'*menyambung hidup*', '*mencabut nyawa*', '*meneteskan air mata*', '*membutuhkan energi*', '*merasa pusing*', and '*mengalami demam*'—reflect the integration of basic physiological need nouns within Indonesian LVCs. These nouns—*hidup* (life), *nyawa* (soul/life force), *air mata* (tears), *energi* (energy), *pusing* (dizziness), and *demam* (fever)—represent biological conditions essential to bodily function and survival. They exhibit strong [+durativity] and consistent [-dynamicity], reflecting their status as ongoing, internally experienced states. Verbs such as *menyambung, membutuhkan,* or *mengalami* function grammatically to predicate these physiological states without encoding change themselves. Even verbs like *mencabut* and *meneteskan,* through more dynamic in form, highlight transitions that culminate in or reference a persistent physical state (*nyawa, air mata*). These constructions remain largely atelic unless externally bounded. Collectively, the data illustrate that basic physiological nouns serve as stative anchors in LVCs, foregrounding bodily continuity and internal states over actionality, with aspectual force primarily noun-driven.

5.2 Eventive nouns

Shifting the focus to eventive nouns within LVCs, the findings propose a three-way classification: indefinite process nouns (§5.2.1), definite process nouns (§5.2.2), and punctual nouns (§5.2.3). This categorization aims to explore potential variations in how they combine with light verbs in Indonesian. Unlike stative nouns, which tend to depict states, eventive nouns represent events or processes. This distinction is reflected in their compatibility with light verbs. Eventive nouns, particularly indefinite process nouns like *main* (lit. play), demonstrate a greater propensity to co-occur with dynamic light verbs such as '*melakukan*' (do), '*membuat*' (make), ‘*memberi*’ (give), and ‘*mengambil*’ (take). These light verbs introduce an agentive dimension, highlighting the ongoing or iterative nature of the indefinite process. Definite process nouns exemplified by ‘*pertemuan*’ (lit. meeting) or ‘*pertandingan*’ (lit. competition) may also participate in LVCs, though with potentially different semantic nuances. Finally, punctual nouns, capturing momentary events like ‘*jatuh*’ (fall) or ‘*teriak*’ (shout), might exhibit distinct combinatorial patterns as well. This classification by their processuality, interact with light verbs to contribute to the overall meaning expressed within LVCs in Indonesian.

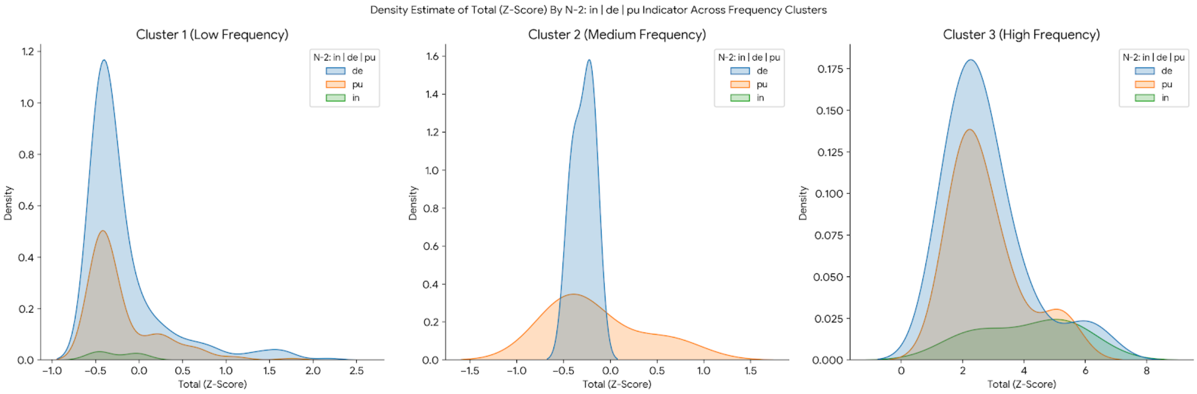


Figure 5.5: Density estimate of indefinite process, definite process, and punctual noun within Indonesian LVCs.

In general, Figure 5.5 delineates the kernel density estimates of total scores across three discrete frequency clusters, categorized by the linguistic aspects of indefinite process (In), definite process (De), and punctual noun (Pu). KDE serves as a non-parametric method to approximate the probability density function of the total scores, thereby offering a smoothed representation on their distribution within each cluster. The abscissa represents the total score, while the ordinate denotes the density indicating the relative likelihood of occurrence for specific score values. Stratification by ‘In’, ‘De’, and ‘Pu’ is visually achieved through distinct shaded regions within each cluster’s subplot, enabling a comparative analysis of how these linguistic categories modulate the distribution of total scores across varying frequency contexts. The overarching aim is to elucidate the interplay between frequency, linguistic categorization, and the resultant distribution of total scores. For instance, a notable observation is that in the low-frequency cluster (Cluster 1), the distribution of punctual nouns (Pu, green) appears to be concentrated at lower total scores compared to indefinite processes (In, blue) and definite processes (De, orange).

Within Cluster 1, designated as the low-frequency cluster, the kernel density estimates reveal nuanced distributional patterns contingent upon the linguistic category. The indefinite process (In) category (blue) exhibits a distribution characterized by a prominent peak around a total score of approximately -0.5, with a noticeable positive skew, indicating a higher density of observations towards the lower end of the total score range, with a tail extending towards higher scores. This suggests that in low-frequency contexts, indefinite processes are predominantly associated with lower total scores, with a smaller proportion of instances yielding higher scores. Conversely, the definite process (De) category (orange) demonstrates a more symmetrical distribution, with a central peak around a total score of approximately 0.0 and slight positive skew, indicating a central tendency towards moderate total scores. The punctual noun (Pu) category (green), while less prominent, shows a distribution skewed towards the lower end, with a peak around -0.5, similar to the indefinite process category. The divergence in distributional characteristics across 'In', 'De', and 'Pu' within Cluster 1 (Low Frequency) underscores the influence of linguistic categorization on total score patterns at lower frequencies.

Cluster 2, representing the medium-frequency cluster, showcases a discernible shift in the distributional patterns of total scores across the linguistic categories. Notably, the indefinite process (In) category is absent from this KDE plot. This absence stems from either a paucity of data points for the 'In' category in this specific cluster, a concentration of those data pints within a narrow range, precluding the generation of a reliable kernel density estimate by Seaborn. The De category exhibits a unimodal distribution centered around a total (Z-score) value of approximately -0.5, with a noticeable spread extending towards higher positive values but with decreasing density. In contrast, the 'Pu' category displays a unimodal distribution with a peak centered around a positive total (Z-score) value of roughly 0.5. This distribution is also relatively spread, indicating a degree of variability in the total scores for this category within the medium frequency cluster. The distinct peaks and partially overlapping distributions suggest that at medium frequencies, there is tendency for the 'Pu' category to be associated with higher total scores compared to the De category, although there is some overlap in their score ranges.

The plot for Cluster 3 (High Frequency) includes three categories: 'In', 'De', and 'Pu'. Both 'De' and 'Pu' categories show unimodal distributions that are shifted towards higher total (Z-score) values compared to the lower frequency clusters. The 'De' category peaks around a total score of approximately 2, while the 'Pu' category peaks slightly higher, around 3. Notably, a third category, 'In' (green), emerges in this high-frequency cluster, exhibiting a bimodal distribution. One peak for 'In' is located around a total score of 3, similar to the peak of Pu, while a second, smaller peak appears around a total score of 6. This suggests that within the high-frequency context, the 'In' category might encompass instances with both moderately high and very high total scores, indicating a more complex pattern compared to the relatively unimodal distributions of 'De' and 'Pu' at this frequency.

Across the frequency clusters, the KDE plots reveal a dynamic relationship between the categories and their total score distributions. At low-frequencies (Cluster 1), 'De' shows a distribution centered around lower values, while 'Pu' is centered slightly higher. Moving to medium-frequencies (Cluster 2), both 'De' and 'Pu' shift towards higher values, maintaining distinct unimodal distributions. The high-frequency cluster (Cluster 3) shows a further shift towards even higher total scores for both 'De' and 'Pu', introducing a bimodal distribution for the 'In' category. This evolution suggests that as the frequency of occurrences increases, the total scores tend to be higher for all categories patterns within specific categories that might be obscured at lower frequencies.

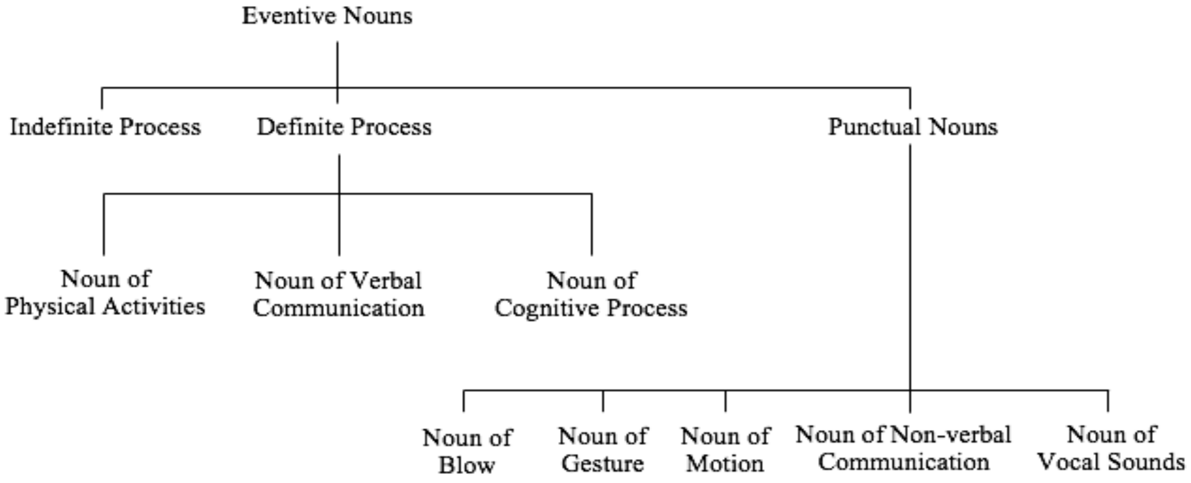


Figure 5.6: Classification of eventive nouns within Indonesian LVCs.

5.2.1 Noun of indefinite process

Remarkably, the analysis of indefinite process nouns within LVCs reveals a limitation, as evidenced by the restricted number of examples presented in Table 5.3. This infrequent occurrence might be attributable to the inherent aspectual properties of indefinite process nouns.

Unlike bounded eventive nouns which represent events with a clear beginning and end, indefinite process nouns, similar to their synthetic verb counterparts, are unbounded and inherently atelic. Atelic essentially signifies the absence of a natural endpoint, characterizing ongoing or repetitive processes. This aspectual incompatibility might explain the limited integration of indefinite process nouns into LVCs. Light verbs themselves often introduce a sense of completion or boundedness. The mismatch between the atelic nature of indefinite process nouns and the potentially telic nature of light verbs might restrict their co-occurrence, hindering the formation of LVCs that represent ongoing or repetitive processes.

Table 5.3: Notable list of the Indonesian LVCs containing indefinite process nouns.

No.	LVC	PMW	No.	LVC	PMW
1.	*membuat integrasi* 'makes an integration'	4.621	6.	*merasakan dampak* 'take influence'	1023.470
2.	*membuat penafian* 'make a disclaimer'	2.612	7.	*mengalami tekanan* 'to come under pressure'	1529.504
3.	*memberikan informasi* 'provides information'	11273.390	8.	*mendapatkan perlindungan* 'to be under protection'	889.214
4.	*menghadapi tantangan* 'be faced with a challenge'	6709.940	9.	*menancapkan hegemoni* 'pursue hegemonism'	12.908
5.	*mengajukan permintaan* 'make a claim'	1066.213	10.	*mempertahankan pengaruh* 'keep some influence'	7.835

In particular, Table 5.3 offers a curated selection of Indonesian LVCs featuring indefinite process nouns, a subcategory of eventive nouns representing events. This compilation sheds light on how Indonesian employs LVCs to express actions, events, or processes that are ongoing, habitual, or lack a clearly defined endpoint. As a detailed description, for '*membuat integrasi*' (makes an integration), this LVC denotes the action of creating or establishing integration. Its low frequency across all corpora (PMW = 4.621) suggests that it might be specialized or domain-specific, potentially appearing more often in technical or academic contexts where the concept of integration is frequently discussed. For '*membuat penafian*' (make a disclaimer), this LVC signifies the act of issuing a disclaimer or denial. Similar to the previous example, its low frequency (PMW = 2.612) might indicate a restricted usage, possibly confined to legal or formal settings. Next, '*memberikan informasi*' (provides information), this LVC represents the action of giving or supplying information. Its remarkably high frequency (PMW = 11273.390) across all corpora underscores its centrality in communication and knowledge dissemination. It emphasizes the strong association between the verb *memberikan* and the noun *informasi* in forming this LVC. As of '*menghadapi tantangan*' (be faced with a challenge), this LVC portrays the experience of encountering or confronting a challenge. Its relatively high frequency (PMW = 6709.940) suggest its common usage in describing situations involving obstacles or difficulties. Additionally, '*mengajukan permintaan*' (make a claim), this LVC denotes the act of putting forth a request or demand. While its frequency is lower than the previous examples (PMW = 1066.213), it still represents a linguistic pattern for expressing claims or requests. For '*merasakan dampak*' (take influence), this LVC conveys the experience of being affected or influenced by something. As of '*mengalami tekanan*' (to come under pressure), this LVC portrays the experience of being subjected to pressure or stress. Its frequency (PMW = 1529.504) are similar to the previous example, suggesting a comparable level of usage in discussing experiences of pressure or hardship.

Besides, '*mendapatkan perlindungan*' (to be under protection), this LVC signifies the state of being shielded or safeguarded. Its low frequency (PMW = 889.214) reflect its importance in specific contexts related to security and safety. For '*menancapkan hegemoni*' (pursue hegemonism), this LVC describes the act of establishing or asserting dominance or control. Its low frequency (PMW = 12.908) suggests its specialized usage, confined to political or social discourse. Lastly, '*mempertahankan pengaruh*' (keep some influence), this LVC denotes the action of maintaining or preserving influence or impact. Similar to the previous example, its low frequency (PMW = 7.835) indicates its restricted usage, potentially within specific domains or contexts of political discourse. It is worthy to note that regarding the semantic diversity, the LVCs in Table 5.3 span a broad semantic contextual-spectrum. This diversity underscores the versatility of Indonesian LVCs in expressing various facets of indefinite processes. Concerning their frequency and collocation strength, the PMW values reveal the varying degrees of usage and collocational strength of these LVCs. High-frequency LVCs with strong PMW values indicate well-established and commonly used expressions, while low-frequency LVCs may represent more specialized or context-dependent constructions.

(5.32)	*Saya*	***mendapatkan***	***perlindungan***	*dari*	*Buddha.*
	I	get	protection	from	Buddha.

'I receive protection from Buddha.'

(5.33)	*Saat*	*ini*	*kami*	*belum*	***merasakan***	***dampak***	*dari*	*El Nino.*
	time	this	we	not yet	feel	impact	from	El Nino

'We haven't felt the impact of El Nino yet.'

(5.34)	*Selama*	*20 hari*	*aku*	***mengalami***	***tekanan***	*psikologis*	*yang*	*dahsyat.*
	for	20 days	I	experience	pressure	psychological	that	devastating

‘For 20 days, I experienced immense psychological pressure.’

(5.35)	*Gereja*	***menghadapi***	***tantangan***	*yang*	*sama.*
	church	face	challenge	that	same

‘The church faces the same challenge.’

(5.36)	*Berpikirlah*	*dua*	*kali*	*untuk*	***memberikan***	***informasi***	*pribadi.*
	think	two	time	for	give	information	personal

‘Think twice before giving out personal information.’

In details, as demonstrated in (5.32) to (5.36), the LVCs *‘mendapatkan perlindungan*’, ‘*merasakan dampak*’, ‘*mengalami tekanan*’, ‘*menghadapi tantangan*’, and *‘memberikan informasi*’ illustrate the behavior of eventive nouns of indefinite process. These nouns—*perlindungan* (protection), *dampak* (impact), *tekanan* (pressure), *tantangan* (challenge), and *informasi* (information)—encode extended, contextually unbounded processes rather than sharply delineated events. Based on inherently features, they exhibit [+dynamicity] and [+durativity], reflecting their gradual unfolding in time. However, they are typically [-telic] and [-bounded], as their completion is not intrinsic but rather discourse-determined. For example, *perlindungan* and *tekanan* imply ongoing states influenced by external agents or conditions, while *informasi* implies a communicative act whose boundaries depend on interactional framing. The verbs paired with these nouns—*mendapatkan, merasakan, menghadapi*—serve to predicate the process without enforcing aspectual closure. These constructions thus embody eventiveness without inherent endpoint, highlighting the indefinite, processual nature of this noun subclass within LVC framework.

5.2.2 Noun of definite process

Delving deeper into the category of eventive nouns within LVCs, we encounter a multifaceted landscape populated by definite process nouns. As illustrated in Figure 5.6, this category encompasses a wide range of subcategories, each contributing distinct semantic nuances to the overall meaning of the constructions. Physical activity-related nouns, exemplified by ‘*berlari*’ (run) or ‘*menari*’ (dance), readily participate in LVCs. These constructions often depict actions undertaken by the experiencer. Communication processes are also well-representated, with nouns like ‘*berbicara*’ (speak) or ‘*menulis*’ (write) forming part of these constructions. The realm of cognitive processes is another significant contributor, with nouns such as '*berpikir*' (think) or '*analisis*' (analyze) readily combining with light verbs. This expansive category highlights the versatility of LVCs in Indonesian, allowing for the expression of a diverse range of eventive processes, encompassing physical actions, communication, and internal cognitive states. In particular, Table 5.4 presents a curated collection of Indonesian LVCs featuring definite process nouns, a subcategory of eventive nouns representing specific and identifiable actions or events. This compilation sheds light on how the Indonesian language employs LVCs to express concrete and distinct processes.

Table 5.4: Notable list of the Indonesian LVCs containing definite process nouns.

No.	LVC	PMW	No.	LVC	PMW
1.	*melakukan perjalanan* 'make a journey'	7864.954	11.	*memberikan komentar* 'give comments'	2747.001
2.	*melakukan kunjungan* 'makes a visit'	7082.371	12.	*memberikan ceramah* 'give a lecture'	744.461
3.	*melakukan perekaman* 'get video'	1007.900	13.	*memberikan pidato* 'give speaking'	401.764
4.	*melakukan pencatatan* 'be on the list'	515.879	14.	*memberikan khotbah* 'give a sermon'	75.591
5.	*melakukan permainan* 'to play'	308.442	15.	*mengambil keputusan* 'take a decision'	15345.272
6.	*mengambil istirahat* 'take a break'	54.144	16.	*melakukan penelitian* 'do research'	6967.754
7.	*membuat pertunjukan* 'make a performance'	49.925	17.	*mengenyam pendidikan* 'to sit in the school'	3985.190
8.	*memberikan apresiasi* 'to commend outstanding ones'	5887.327	18.	*memberikan pelajaran* 'give a lesson'	1023.671
9.	*memberikan penjelasan* 'give explanation'	5880.245	19.	*membentuk pemikiran* 'do think'	23.607
10.	*memberikan sambutan* 'give a speech'	3770.320	20.	*mengambil terobosan* 'take the iniciative'	13.109

5.2.2.1 Noun of physical activities

First subdivision of the definite process noun within Indonesian LVCs is noun of physical activities (Pa). Based on Table 5.4 (*see* No. 1-7), this category encompasses LVCs that denote specific physical actions or activities, often involving movement or change in location. This type of eventive noun encompasses all those nouns that signify physical activity undertaken by any entity capable of movement. Physical activity, in this context, is broadly defined as any exertion or action that primarily involves bodily engagement and does not necessitate significant mental or cognitive processing. This encompasses a wide spectrum of activities, from basic movements like walking or running to complex actions requiring skill and coordination. In Indonesian LVCs, these eventive nouns can represent an immense assortment of physical activities, capturing the nature of physically engagement in the world. The interplay between these eventive nouns and light verbs further enriches the expressive potential, allowing for nuanced descriptions of various aspects of physical activity, such as its intensity, duration, and purpose.

(5.37) *Anak* *ini* ***melakukan*** ***perjalanan*** *panjang.*
child this conduct journey long
‘This child went on a log journey.’

(5.38) *Mereka* *juga* ***melakukan*** ***kunjungan*** *sosial.*
they also conduct visit social
‘They also made a social visit.’

(5.39) *Ia* *tak* *pernah* ***melakukan*** ***permainan*** *aneh* *lagi.*
he/she not ever conduct game strange again
‘He/she never played strange games again.’

(5.40) *Peneliti* ***melakukan*** ***perekaman*** *audio* *video* *belajar.*
researcher conduct recording audio video learning.
‘The researcher recorded the learning session on audio and video.’

(5.41) *Mereka* *juga* ***melakukan*** ***pertunjukan*** *di* *Amerika Serikat.*
they also conduct performance in United States
‘They also gave a performance in the United States.’

As showed in Table 5.4, ‘*melakukan perjalanan*’ (make a journey), this LVC, with the substantial PMW value of 7864.954, clearly indicates its widespread usage in describing the act of traveling. Also, ‘*melakukan kunjungan*’ (makes a visit), another frequent LVC (PMW = 7082.371), signifying the act of visiting someone or a place, further highlighting the prominence of physical activities within this category. Next, ‘*melakukan perekaman*’ (get video), ‘*melakukan pencatatan*’ (be on the list), ‘*melakukan permainan* (to play), these LVCs, although less frequent, represent various physical activities involving recording, listing, and playing, respectively. Their PMW values suggest their more specific contextual usage. The range of PMW values suggests that the specificity and commonality of the physical action influence the overall usage of the LVC.

Furthermore, as showed in (5.37) to (5.41) the LVCs '*melakukan perjalanan*', '*melakukan kunjungan*', '*melakukan permainan*', '*melakukan perekaman*', and '*melakukan pertunjukan*' exemplify the behavior of physical activity nouns within. These nouns—*perjalanan* (journey), *kunjungan* (visit), *permainan* (game), *perekaman* (recording), and *pertunjukan* (performance)—represent bounded, goal-oriented physical actions. They exhibit a strong profile of [+dynamicity], [+durativity], [+telicity], and often [+boundedness], as these processes unfold over time with identifiable endpoints. The telicity is lexicalized in the nominal domain—e.g., *kunjungan* presupposes arrival and interaction, while *perekaman* entails technological mediation of a captured event. The verb *melakukan* in all five constructions contributes minimal aktionsart value, functioning as a light verb to license predication. Its neutrality enables the noun to carry the full eventive load. These LVCs demonstrate the temporal richness and structural clarity of definite physical activities, where the noun defines the dynamic contour, bounded trajectory, and goal-completion structure of the LVC as a whole.

5.2.2.2 Noun of verbal communication

Second subdivision of the definite process noun within Indonesian LVCs is noun of verbal communication (Vc). Based on Table 5.4 (*see* No. 8-13), This category comprises LVCs that represent specific acts of verbal communication. This type of eventive noun encompasses all those nouns that represent instances of verbal communication, whether they occur in spoken or written form. Verbal communication, fundamentally, is defined as the interactive process between language users who employ linguistics tools, such as words, phrases, and sentences, as a means of exchanging information, ideas, and emotions. In the context of Indonesian LVCs, these eventive nouns can signify a diverse assortment of verbal communication events, including acts of speaking (*bicara*), writing (*tulis*), reading (*baca*), or even listening (*dengar*). The interplay between these eventive nouns and light verbs further enriches their expressive potential.

(5.42) *Kami* ***memberikan*** ***apresiasi*** *yang* *tinggi* *untuk* *itu.*
we give appreciation that high for that
'We give high appreciation for that.'

(5.43) *Perawat* *perlu* ***memberikan*** ***penjelasan*** *dan* *penjelasan.*
nurse need give explanation and counseling
'The nurse needs to provide explanation and counseling.'

(5.44)	*Saya*	*tidak*	*perlu*	***lagi***	***memberikan***	*sambutan.*		
	I	not	need	anymore	give	welcome speech		
	‘I don’t need to give a welcome speech anymore.’							
(5.45)	*Guru*	*lebih*	*banyak*	***memberikan***	***komentar***	*dan*	*umpan balik.*	
	teacher	more	much	give	comment	and	feedback	
	‘The teacher gives more comments and feedback.’							
(5.46)	*Beliau*	*juga*	***memberikan***	***khotbah***	*Dharma.*			
	he/she	aslo	give	sermon	Dharma.			
	‘He also gives a Dharma sermon.’							

For instance, ‘*memberikan apresiasi*’ (to commend outstanding ones), ‘*memberikan penjelasan*’ (give explanation), ‘*memberikan sambutan*’ (give a speech), these LVCs, with relatively moderate PMW, all utilize the verb ‘*memberikan*’ (to give) to express different forms of verbal communication, highlighting its significance in this domain. Moreover, ‘*memberikan komentar*’ (give comments), ‘*memberikan ceramah*’ (give a lecture), ‘*memberikan pidato*’ (give speaking), these LVCs further illustrate various forms of verbal expression, with decreasing frequencies suggesting their more specialized or formal contexts of usage. The inclusion of ‘*memberikan khotbah*’ (give a sermon), although infrequent, adds a religious dimension to the types of verbal communication represented. As the implications for this sub-category, the prevalence of ‘*memberikan*’ in this sub-domain emphasizes its role in framing acts of communication as the giving or bestowing of verbal content. The range of PMW values reflects the varying formality and specificity of the communicative acts represented.

Moreover, the LVCs ‘*memberikan apresiasi*’, ‘*memberikan penjelasan*’, ‘*memberikan sambutan*’, ‘*memberikan komentar*’, and ‘*memberikan khotbah*’ intensely exemplify the subclass of verbal communication nouns. These nouns—*apresiasi, penjelasan, sambutan, komentar,* and *khutbah*—encode structured communicative acts that are purpose-driven, temporally bounded, and lexically telic. These nouns exhibit strong [+dynamicity], [+durativity], [+telicity], and frequently [+boundedness]. For instance, *penjelasan* and *sambutan* denote finite speech events with clear beginnings and endings, while *khotbah* presumes a performative ritual. The verb *memberikan* contributes minimal aspectual weight, serving instead to license the nominal predicate in a structurally schematic way. Its lightness enables the noun to determine the eventive contour of the construction. Collectively, these LVCs demonstrate how verbal communication nouns structure LVCs as bounded, goal-oriented discourse acts, aligning with the semantic properties of the definite process category and reinforcing the primacy of the noun in shaping eventiveness within Indonesian predicate constructions.

5.2.2.3 Noun of cognitive process

Third subdivision of the definite process noun within Indonesian LVCs is of the definite process noun within Indonesian LVCs is noun of cognitive process (Co) (*see* No. 14-20 of Table 5.4). This category includes LVCs that denote specific mental or cognitive processes. This type of eventive noun is intrinsically linked to all those nouns that represent the multifaceted realm of cognitive processes. Cognitive processes, in essence, are defined as those fundamental activities undertaken by language users that are centered on the intricate workings of the mind. These processes encompass a broad spectrum of mental operations, including perception, attention, memory, reasoning, decision-making, and problem-solving. In the context of Indonesian LVCs, these eventive nouns can encapsulate a classified assortment of cognitive events, such as acts of thinking (*pikir*), remembering (*ingat*), and understanding (*paham*). The interaction between these eventive nouns and light verbs adds a layer of dynamism, allowing for the expression of not only static cognitive states but also the unfolding processes of mental activity.

(5.47) *Mereka lalu* ***mengambil keputusan*** *untuk mengirim utusan.*
they then take decision for send messenger
'They then made a decision to send a messenger.'

(5.48) *Dia banyak* ***melakukan penelitian*** *tentang obat.*
he/she many conduct research about medicine
'He conducted a lot of research about medicine.'

(5.49) *Ia* ***mengenyam pendidikan*** *tinggi hingga ke Cambridge.*
he/she undergo education high until to Cambridge
'He pursued higher education all the way to Cambridge.'

(5.50) *Buku ini* ***memberikan pelajaran*** *berharga bagi saya.*
book this give lesson valuable for me
'This book provides a valuable lesson for me.'

(5.51) *Inti filsafat adalah* ***membentuk pemikiran.***
core philosophy is form thinking
'The essence of philosophy is to shape thought.'

For instance, '*mengambil keputusan*' (take a decision), this LVC, with the highest PMW value (15345.272) in the entire Table 5.4, signifies the crucial cognitive process of decision-making. Also, '*melakukan penelitian*' (do research), '*mengenyam pendidikan*' (to sit in the school), these frequent LVCs highlight the importance of research and education as cognitive processes. Moreover, '*memberikan pelajaran*' (give a lesson), '*membentuk pemikiran*' (do think), '*mengambil terobosan*' (take the initiative), these LVCs, with decreasing frequencies, represent various cognitive actions, ranging from instruction to independent thought and innovation. It can be implied that the diversity of verbs in this sub-domain reflects the multifaceted nature of cognitive processes. Particularly, the LVCs '*mengambil keputusan*', '*melakukan penelitian*', '*mengenyam pendidikan*', '*memberikan pelajaran*', and '*membentuk pemikiran*' represent the essence of cognitive process derived from the noun elements. These nouns—*keputusan* (decision), *penelitian* (research), *pendidikan* (education), *pelajaran* (lesson), and *pemikiran* (thought)—encode mental activities that unfold over time and lead to definable outcomes. They exhibit strong [+durativity], [+dynamicity], and context-sensitive [+telicity], as in *keputusan* and *pelajaran,* which imply resolution or instructional delivery. Their [+boundedness] varies, with some process—like *pendidikan* and *penelitian*—extending over long durations, yet still directed toward specific cognitive achievements. Verb such as *mengambil, melakukan,* or *memberikan* serve as a light grammatical role, allowing the noun to determine the temporal structure and conceptual scope of the event. These LVCs demonstrate how cognitive process nouns encode intellectual change and development, and function as the primary carriers of eventiveness, defining the LVC's temporal and goal-oriented contours.

5.2.3 Punctual nouns

The analysis of eventive nouns within LVCs extends to the category of punctual nouns, which capture momentary events or actions. Punctual nouns encompass a surprisingly diverse range of semantic domains. Actions like blows, exemplified by '*pukul*' (hit) or '*tendang*' (kick), readily participate in LVCs, often depicting a single, forceful act. Expressions of love, such as '*cium*' (kiss) or '*peluk*' (hug), also find a place within this category, adding a layer of emotional nuance to LVCs. Interestingly, motion nouns, like '*lompat*' (jump) or '*jatuh*' (fall), can also function as punctual nouns within LVCs, highlighting a single instance of movement. The realm of communication is further explored through nouns related to verbal or non-verbal communication, such as '*teriak*' (shout) or '*lambai*' (wave). Finally, vocal sound nouns, exemplified by '*batuk*' (cough) or '*tawa*' (laugh), contribute to the tapestry of punctual events expressed through LVCs. This diverse range of punctual nouns underscores the flexibility of LVCs in Indonesian, allowing for the representation of a multitude of momentary actions and events.

Table 5.5: Notable list of the Indonesian LVCs containing punctual nouns.

No.	LVC	PMW	No.	LVC	PMW
1.	*memadamkan kebakaran* 'take fire'	408.343	11.	*melakukan peletakan* 'stand in place'	186.693
2.	*menghadapi letusan* 'be about to explode'	4.018	12.	*melakukan pergantian* 'be about to be handed over'	546.567
3.	*mencium aroma* 'have a smell perceived'	821.659	13.	*memberikan dukungan* 'receives support'	5369.640
4.	*masuk angin* 'take a cold'	1704.997	14.	*memberikan perhatian* 'give one's mind to'	4651.900
5.	*mengambil langkah* 'takes a step'	7101.056	15.	*memberikan perlindungan* 'take into custody'	4638.489
6.	*melepaskan tendangan* 'give a blow'	2656.140	16.	*memberikan pertolongan* 'give a hand'	1426.540
7.	*menarik nafas* 'take a sigh'	1429.302	17.	*mendengarkan musik* 'to listen to music'	3488.347
8.	*memberikan sentuhan* 'give a touch'	746.671	18.	*mengeluarkan suara* 'make a noise'	1574.859
9.	*mengambil tindakan* 'take action'	6144.236	19.	*menyampaikan salam* 'give regards'	562.991
10.	*melakukan pendaratan* 'make landing'	430.343	20.	*memberikan teguran* 'give reprimand'	540.540

Specifically, Table 5.5 presents a curated selection of Indonesian LVCs that incorporate punctual nouns, a category signifying momentary events or actions. This collection offers insights into how Indonesian employs LVCs to express discrete and time-bound occurrences. The table further classifies these LVCs into five sub-domains, each accompanied by frequency metrics from four corpora under examination, allowing for a nuanced examination of their usage patterns and contextual relevance.

5.2.3.1 Noun of blow

First subdivision of the punctual noun within LVCs is noun of blow (Bl) (*see* No. 1-4 of Table 5.5). This category encompasses LVCs that represent actions involving a sudden impact or forceful contact. For instance, '*memadamkan kebakaran*' (take fire), this LVC depicts the act of extinguishing a fire, a forceful action with immediate consequences. Its relatively low moderate frequency (PMW value of 408.343) suggest its specific context of use. Also, '*menghadapi letusan*' (be about to explode), this LVC denotes the experience of facing an imminent explosion, a sudden and impactful event. Its PMW (4.018) in the corpora indicates its infrequent usage, possibly limited to specific genres or contexts. Besides, as of '*mencium aroma*' (have a smell perceived), while seemingly a sensory experience, this LVC also implies a momentary perception of an odor. Its moderate frequency (PMW value of 821.659) suggest its broader applicability in describing various situations. Lastly, '*masuk angin*' (take a cold), this idiomatic LVC represents the sudden onset of a cold or flu-like symptoms. Its relatively high PMW value (1704.997) highlight its common usage in everyday language.

(5.52) *Masyarakat memahami ancaman* ***menghadapi letusan*** *gunung Merapi.*
society understand threat facing eruption mountain Merapi
'The community understands the threat of facing the eruption of Mount Merapi.'

(5.53) *Hujan lebat* ***memadamkan kebakaran*** *yang telah terjadi.*
rain heavy extinguish fire that already happen
'The heavy rain extinguished the fire that had occurred.'

(5.54) *Dalam pikiranku dia* ***masuk angin*** *dan kelelahan.*
in my-mind he/she enter wind and exhaustion
'I think he's feeling under the wheater and exhausted.'

In details, as illustrated in (5.52) to (5.54), the LVCs 'menghadapi letusan', 'memadamkan kebakaran', and 'masuk angin' illustrate the temporal behavior of nouns of blow, characterized by sudden, forceful phenomena. In 'menghadapi letusan' (to face an eruption), letusan (eruption) encodes a highly bounded, telic event marked by [+dynamicity], [+telicity], and [+boundedness], while the verb menghadapi introduces an agentive stance toward a punctual natural occurrence. Similarly, 'memadamkan kebakaran' (to extinguish fire) centers on kebakaran (fire), another forceful event with a clear start and end. Here, memadamkan signals resolution, reinforcing the noun's [+telicity] and [+boundedness]. By contrast, 'masuk angin' (to catch cold) is metaphorical and idiomatic yet preserves the aspectual structure of a sudden-onset state—angin functioning as a nominal trigger for bodily disruption, best described as [+telic] and [+dynamicity]. Collectively, these LVCs demonstrate how nouns in the blow category impose punctual, event-driven interpretations, often centered on disruption, force, and immediate transformation.

5.2.3.2 Noun of gesture

Second subdivision of the punctual noun within LVCs is noun of gesture (Ge) (*see* No. 5-8 of Table 5.5). This particular type of eventive noun holds a distinct position due to its specific reference to all those nouns that are intertwined with the utilization of body gestures or postures within the realm of non-verbal communication. Gestures, in this context, are defined as the deliberate employment of various human body parts, such as hands, arms, head, or facial muscles, to serve as meaningful signals in interpersonal exchanges. These signals can convey a wide spectrum of information, from emotions and attitudes to intentions and requests, often supplementing or even replacing verbal communication. In Indonesian LVCs, these eventive nouns can encapsulate a repertoire of gestural expressions, from subtle nods and winks to elaborate hand movements and full-body displays. The interplay between these eventive nouns and light verbs supplementary augments their easy-to-read command, allowing for the representation of not just stagnant gesture but also the active recitation of gestural communication in real time interaction.

(5.55) *Kita bisa* ***memberikan sentuhan*** *personal.*
we can give touch personal
'We can give a personal touch.'

(5.56) *Saya* ***menarik nafas*** *panjang agar tambah yakin.*
I pull breath long so that add sure
'I took deep breath to feel more confident.'

(5.57) *Firman Utina* ***melepaskan tendangan*** *keras dari jarak jauh.*
Firman Utina release kick hard from distance far
'Firman Utina unleashed a powerful shot from long range.'

For instance, '*mengambil langkah*' (takes a step), this high-frequency LVC with a PMW value of 7101.056 represents the fundamental act of taking a step, often used metaphorically to indicate progress or decision-making. Following, '*melepaskan tendangan*' (give a blow), this LVC signifies the action of delivering a kick or a blow, highlighting a forceful physical gesture. Then, '*menarik nafas*' (take a sigh), similar to its counterpart in this category, this LVC denotes the act of sighing, a brief but meaningful physical expression. Also, '*memberikan sentuhan*' (give a touch), this LVC represents the act of touching, a gesture that can convey various emotions or intentions. In all, it can be assumed that the LVCs in this category demonstrate the linguistic representation of physical gestures, often imbued with communicative or symbolic meaning, and the varying PMW values highlight the diverse range of gestures and their contextual relevance.

Moreover, as exemplified in (5.55) to (5.57), the LVCs ‘*memberikan sentuhan*’, ‘*menarik nafas*’, and ‘*melepaskan tendangan*’ represent the subclass of gesture nouns within the punctual eventive noun category. These nouns—*sentuhan* (touch), *nafas* (breath), and *tendangan* (kick)—refer to bodily actions that occur as discrete, temporally bounded gestures. They are best characterized as [+dynamicity], [+telicity], and [+boundedness], often with brief or momentary duration. *Sentuhan,* in ‘*memberikan sentuhan*’*,* connotes a deliberate yet ephemeral act, whose meaningfulness is often discursively amplified despite its physical brevity. ‘*Menarik nafas*’ encodes a biologically grounded yet symbolically rich action—durative in execution but telic in framing. In contrast, ‘*melepaskan tendangan*’ is a prototypical gesture noun LVC: the noun *tendangan* conveys a complete kinetic event with clear trajectory and endpoint. In all three, the verb serves to trigger the gestural frame, while the noun encapsulates the full aspectual structure. These constructions underscore how gesture nouns in LVCs function as eventive anchors for physically situated, symbolically potent, and temporally punctual expressions.

5.2.3.3 Noun of motion

Third subdivision of the punctual noun within LVCs is noun of motion (Mo) (*see* No. 9-12 of Table 5.5). This type of eventive noun encompasses all those nouns that represent the concept of movement. Movement, in its essence, denotes the displacement of an entity from one spatial position to another. This displacement, importantly, unfolds within the dimensions of both time and space, implying a process with duration and direction. In the context of Indonesian LVCs, these eventive nouns can encapsulate a vast array of motion events. For instance, ‘*mengambil tindakan*’ (take action), this frequent LVC (PMW = 6144.236) signifies the act of taking initiative or performing a specific action. Also, ‘*melakukan pendaratan*’ (make landing), ‘*melakukan peletakan*’ (stand in place), and ‘*melakukan pergantian*’ (be about to be handed over), these LVCs, with varying frequencies, describe specific movements or positional changes.

(5.58)	*T 538*	*tergelincir*	*saat*	***melakukan***	***pendaratan***	*di*	*bandara*	*Adisumarmo.*
	T 538	slipped	while	doing	landing	at	airport	Adisumarmo

‘T 538 slipped while landing at Adisumarmo Airport.’

(5.59) *Gubernur Gorontalo* ***melakukan peletakan*** *batu pertama.*
governor Gorontalo doing laying stone first
'The Governor of Gorontalo is laying the first stone.'

(5.60) *Bangsa Indonesia* ***melakukan pergantian*** *paradigma perjuangan.*
nation Indonesia did changing paradigm struggle
'The Indonesian people underwent a paradigm shift in their struggle.'

In details, as demonstrated in (5.58) to (5.60), the LVCs '*melakukan pendaratan*', '*melakukan peletakan*', and '*melakukan pergantian*' reflect the use of motion-oriented nominal cores that encode discrete shifts or transitions. As nouns of motion under the punctual eventive noun class, *pendaratan* (landing), *peletakan* (laying), and *perganting* (shift) denote events with definable initiation and conclusion points, thus carrying strong [+telicity] and [+boundedness]. These actions also demonstrate [+dynamicity], as they involve a directed movement or repositioning—either physical (*pendaratan, peletakan*) or conceptual (*pergantian* as in *pergantian paradigma*). Although the verb *melakukan* is aspectually neutral, it functions to license the event noun syntactically, allowing the noun to dictate the temporal and dynamic frame of the construction. In all cases, the event is construed as singular and punctuated, even when metaphorical, as with *pergantian paradigma*. These LVCs demonstrate how motion nouns operate as the primary carriers of aspectual and eventive structure, shaping the LVC as a bounded, one-time occurrence with inherent directionality.

5.2.3.4 Noun of non-verbal communication

Fourth subdivision of the punctual noun within LVCs is noun of non-verbal communication (Nc) (*see* No. 13-16 of Table 5.5). This type of eventive noun encompasses all types of nouns that are associated with meaning in non-verbal of communication. Non-verbal communication, in this context, refers to a broad spectrum of interaction strategies that rely on the transmission of messages through channels other than spoken or written language. These channels can encompass a variety of modalities, including facial expressions, body language, touch, silence, proxemics (the use of space), and even paralinguistic features such as tone of voice and intonation. Crucially, eventive nouns in Indonesian LVCs can capture the subtle nuances of non-verbal communication, reflecting the wall-hanging of human expression beyond words. For instances, '*memberikan dukungan*' (receives support), '*memberikan perhatian*' (give one's mind to), '*memberikan perlindungan*' (take into custody), '*memberikan pertolongan*' (give a hand), these LVCs, all utilizing the verb '*memberikan*' (to give), signify various forms of non-verbal support, attention, protection, and assistance.

(5.61) *Orang tua* **memberikan perhatian** *lebih terhadap anak bungsu.*
parents give attention more towards child youngest
'Parents give more attention to the youngest child.'

(5.62) *Para dokter* **memberikan dukungan** *yang amat besar.*
doctors give support which very big
'The doctors give very big support.'

(5.63) *Mereka cepat* **memberikan pertolongan** *kepada orang-orang.*
they quicly give help to people
'They quickly offer help to the people.'

(5.64) *KBRI* **memberikan perlindungan** *hukum.*
Indonesian embassy give protection legal
'The Indonesian Embassy provides legal protection.'

In particular, as demonstrated in (5.61) to (5.64) the LVCs '*memberikan perhatian*', '*memberikan dukungan*', '*memberikan pertolongan*', and '*memberikan perlindungan*' demonstrate the use of verbal communication nouns that function within the framework of punctual eventive nouns, particularly those that encode discrete acts of support, aid, or intervention. Nouns such as *perhatian* (attention), *dukungan* (support), *pertolongan* (help), and *perlindungan* (protection) exhibit varying degrees of [+dynamicity] and [+telicity], often performed as one-time communicative or institutionalized acts. While these nouns imply processes that could be durative in nature, within the construction they are often interpreted as bounded communicative gestures, especially in contexts of aid or formal interaction. The light verb *memberikan* activates the noun without contributing additional temporal contour, allowing the noun to govern the aspectual frame. These LVCs straddle the line between verbal expression and institutional action, emphasizing that nouns of verbal communication can function as punctual, telic events, particularly when socially or performatively marked in Indonesian context.

5.2.3.5 Noun of vocal sounds

Fifth subdivision of the punctual noun within LVCs is noun of vocal sounds (Vs) (*see* No. 17-20 of Table 5.5). This is an eventive noun which has a reference meaning around sounds or sound associations. In this context, sound is defined as any sound that can be heard by the human senses. Crucially, eventive nouns within LVCs can encompass a wider range of auditory experiences than just those perceptible through human hearing. This can include sounds from natural world (e.g., *angin* 'wind,' *guntur* 'thunder'), human-produced sounds (e.g., *musik* 'music', *tepuk* 'clap', *teriak* 'shout'), and even non-acoustic phenomena that are metaphorically construed as sounds (e.g., *kilat* 'lighting', *getar* 'vibration'). The specific sound association evoked by an eventive noun can be modulated by the verb it combines with in the LVC. This interaction between the eventive noun and the light verb allows for a contextualized expression of various auditory events and experiences within Indonesian. For instances, *mendengarkan musik* ('to listen to music), *mengelurakan suara* (make a noise), *menyampaikan salam* (give regards), *memberikan teguran* (give reprimand), these LVCs showcase different types of vocal sounds, ranging from enjoying music to expressing greetings or reprimands.

(5.65) *Anda juga dapat* ***mendengarkan musik*** *langsung dari hasil ponsel.*
You also can listen music directly from results cellphone
'You can also listen to music directly from your phone's output.'

(5.66) *Dia* ***mengeluarkan suara*** *melengking panjang.*
he/she take out sound shrill long
'He let out a long, shrill sound.'

(5.67) *Saat ini kami ingin* ***menyampaikan salam*** *dan rasa terima kasih.*
time this we want convey greetings and feeling thank you
'At this time, we want to express our greetings and gratitude.'

(5.68) *Sebaiknya memang guru* ***memberikan teguran*** *lisan secara individual.*
It's best indeed teacher give reprimand oral in-a-way individual
'It is best if teacher gives an oral reprimand individually.'

Furthermore, as exemplified in (5.65) to (5.68), the LVCs '*mendengarkan musik*', '*mengeluarkan suara*', '*menyampaikan pesan*', and '*memberikan teguran*' exemplify the role of vocal sound nouns in Indonesian LVCs as punctual, communicative events. Nouns like *musik* (music), *suara* (sound), *salam* (greeting), and *teguran* (reprimand) encode acoustic or spoken phenomena that are temporally bounded and pragmatically discrete. These nouns exhibit consistent [+dynamicity], [+telicity], and [+boundedness], though *musik* may also display [+durativity] when interpreted as a continuous auditory experience. In '*mengeluarkan suara*' and '*menyampaikan salam*', the nouns represent completed vocal gestures, while '*memberikan teguran*' involvels a directive act that concludes upon delivery. The verbs used—*mendengarkan, mengeluarkan, menyampaikan,* and *memberikan*—function as grammatical triggers, enabling the noun to impose the aspectual profile of the construction. These LVCs show how vocal expressions are encoded as bounded, eventive episodes, reinforcing the punctuality and communicative force of the *noun of vocal sounds* subclass in Indonesian.

5.3 Discussion

This discussion aims to synthesize the salient distinguishing features of stative (§5.3.1) and eventive (§5.3.2) nouns within Indonesian LVCs. This analysis is grounded in both a corpus-based examination of Indonesian texts and pertinent theoretical linguistic frameworks. The objective is to offer a comprehensive overview of the characteristics that differentiate stative nouns from other noun types within this construction, and vice versa. We will leverage the insights gleaned from the six parameters and LVCs frequency and distribution analyses (Chapter 3) to illuminate the unique contributions of both stative and eventive nouns to the overall properties of Indonesian LVCs. The principle of meaning compositionality, which posits that the meaning of a complex expression is derived from the meanings of its constituent parts and their mode of combination, will be central to this exploration. Further, the frequency and distributional patterns of these nouns will serve as empirical evidence to support our theoretical understanding.

5.3.1 Features of stative nouns within Indonesian LVCs

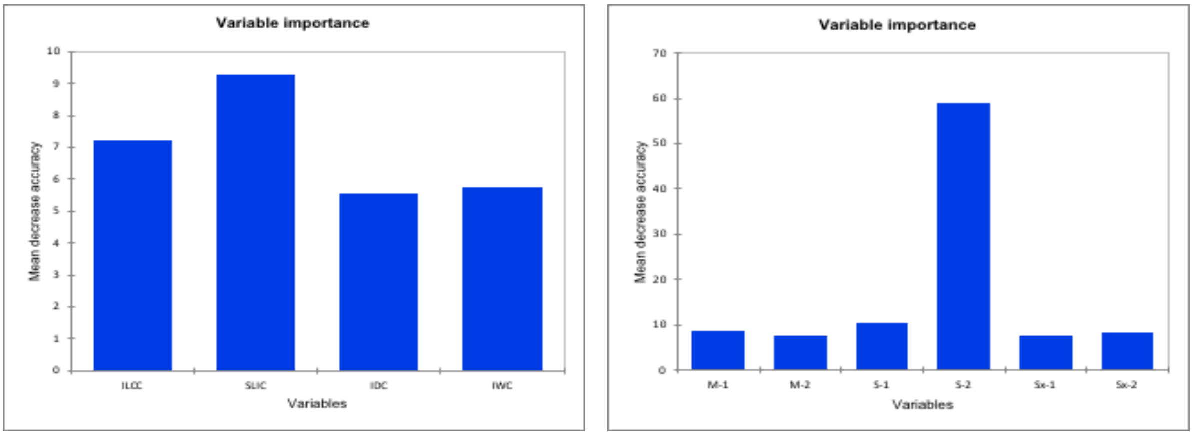
A challenging step within the present study is to identify and summarize the distinctive features of the stative nouns within Indonesian LVCs datasets. The main reason is that we cannot solely rely on a single theoretical understanding to comprehensively explain these features. The inherent complexity and variability of stative nouns within LVCs necessitate a multi-faceted approach that transcends traditional linguistic analysis. Hence, the application of machine learning algorithms offers a promising avenue to enhance our capacity to identify and summarize the salient features that differentiate stative nouns from their eventive counterparts. By leveraging computational techniques to analyze large-scale corpus data, we can uncover patterns and correlations that might elude manual inspection. The subsequent discussion will present a set of proposed distinctive features of stative nouns, meticulously curated based on evidence-based and theoretically sound linguistic principles. These features will serve as a valuable heuristic for distinguishing stative nouns within the context of Indonesian LVCs, contributing to a more refined understanding of their semantic and syntactic behavior.

Table 5.6: Variable importance from a Random Forest Model for predicting *stative* and *eventive noun status* (Part A: Corpora; Part B: Linguistic Principles).

PART A				PART B			
Variables	**Stative nouns**	**Eventive nouns**	**Mean decrease accuracy**	**Variables**	**Stative nouns**	**Eventive nouns**	**Mean decrease accuracy**
ILCC	-1.241	7.179	7.174	M-1	9.657	2.147	8.624
SLIC	8.171	3.941	**9.235**	M-2	13.761	-3.105	7.420
IDC	-0.144	4.863	5.525	S-1	4.417	9.502	10.133
IWC	7.962	-2.257	5.726	S-2	41.440	47.359	**58.891**
				Sx-1	-6.239	11.785	7.378
				Sx-2	-4.769	11.825	8.141

Empirically, quantitative variables such as the frequency size within all corpora under examination have emerged as a substantial factor influencing the selection of noun types within Indonesian LVCs. Table 5.6 (Part A) provides valued insights into the relative importance of different linguistic corpora in predicting the functional distinction between stative and eventive nouns within Indonesian LVCs. Several key observations are as follows. First, ILCC's Dominance. ILCC emerges as the most influential predictor, with the highest mean decrease in accuracy (13.040). Its substantial impact on both ST (4.127) and EV (8.040) suggests that the linguistic patterns and noun usage within ILCC are highly informative in differentiating between stative and eventive nouns within LVCs. Second, IDC's importance for eventive nouns. The IDC corpus also demonstrates considerable importance, with a mean decrease in accuracy of 6.897. While its impact on predicting stative nouns (ST: 0.386) is minimal, its influence on predicting eventive nouns (EV: 7.439) is notable. This suggests that IDC might be particularly rich in examples or linguistic contexts that characterize eventive nouns within LVCs. Third, SLIC and IWC. These two corpora exhibit lower overall importance compared to ILCC and IDC. SLIC shows a moderate impact on both noun classes (ST: 2.321, EV: 3.124), while IWC's influence is slightly weaker (ST: 1.822, EV: 2.457).

Therefore, there are several linguistic implications according to the observed data in Table 5.6 (Part A). Regarding the corpus specificity, the varying importance of the corpora suggests that the linguistic features and contexts captured in each corpus differ in their relevance to distinguishing between stative and eventive nouns within Indonesian LVCs. In respect to the ILCC's prominence, the high importance of ILCC underscores its value as a resource for understanding the subtle distinctions between these noun types. Additionally, the strong association of IDC is not with stative nouns. However, while less influential than ILCC, SLIC and IWC still contribute to the prediction of noun types, suggesting that they capture some linguistic features relevant to the distinction.



(a) (b)

Figure 5.7: Variable importance assessment for stative or eventive noun prediction: (a) corpora-based frequency (quantitative) and (b) linguistic principles (qualitative).

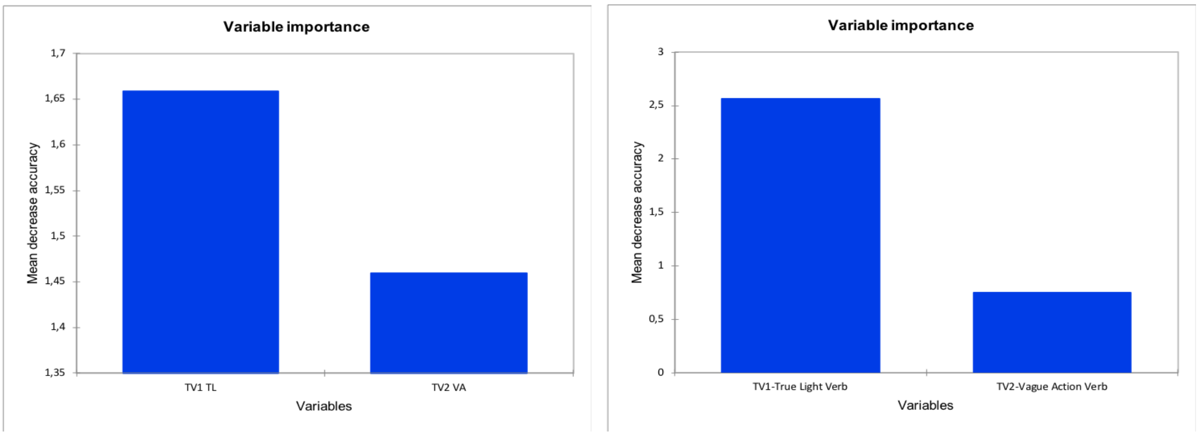
Moreover, regarding the linguistic principles' variables for stative noun prediction within Indonesian LVCs context, Table 5.6 (Part B) presents the results of a variable importance analysis, generated using a Random Forest model. The focus here is on discerning the predictive power of various theoretical linguistic principles in classifying nouns within LVCs as either stative (denoting a state or condition) or eventive (representing an action or event). According to the metric, several key observations can be resulted. First, Semantic Principle of "Noun Core Proposition" (S-2). This principle emerges as the most influential predictor, with the highest mean decrease in accuracy (58.891). Its substantial impact on both St (41.440) and Ev (47.359) classes underscores its critical role in differentiating stative and eventive nouns within LVCs.

It implies that the extent to which the noun carries the core meaning of the LVC is a pivotal factor in determining the noun's nature.

Second, Syntactic Principle of "Transitivity" (Sx-1). While exhibiting a lower overall mean decrease in accuracy (7.378), this principle shows an intriguing pattern. It has a negative value for St (-6.239) but a positive value for Ev (11.785). This suggests that disrupting the transitivity principle actually improves the model's ability to predict stative nouns, while hindering its prediction of eventive nouns. This counter-intuitive result warrants further investigation into the relationship between transitivity and the stative noun category. Lastly, the remaining principles demonstrate varying degrees of importance. The morphological principle of "abstract" vs. "concrete" verbs (M-2) and the semantic principle of the availability of a "synonymous counterpart" (S-1) display moderate predictive power, while the morphological principle of "base" vs. "affixed" verbs (M-1) and the syntactic principle of "valency" (Sx-2) exhibit comparatively lower importance.

Therefore, at least, there are three notable implications from this analysis. In term of semantic prominence, the dominance of the "noun core proposition" principle (S-2) highlights the significance of semantic considerations in classifying nouns within LVCs. This suggests that the semantic role of the noun, particularly its contribution to the core meaning of the construction, is a key factor in determining its stative or eventive nature. As of syntactic complexity, the complex behavior of the transitivity principle (Sx-1) suggests that the relationship between transitivity and the stative/eventive distinction is not straightforward. Further analysis is needed to unravel the nuanced interplay between syntactic structure and noun classification. Lastly, for the morphological and other syntactic considerations, while less influential than semantic and certain syntactic factors, morphological features and valency still contribute to the classification process.

Furthermore, in order to discuss the features of stative nouns, it is also important to measure the correlation between TYPE_VERB and the subsequent division of stative nouns. Figure 5.11 is a representation of the analysis results amongst these types of correlations. Based on calculation, Table 5.7 (Part A) presents the findings of a variable importance analysis derived from a Random Forest model. This model aims to predict the presence of state nouns within Indonesian LVCs based on the TYPE_VERB employed. The analysis specifically examines two categories of verbs: True Light Verbs (TV1 TL) and Vague Action Verbs (TV2 VA), assessing their relative importance in predicting the occurrence of two types of stative nouns: Physiological states (Ph) and Psychological states (Ps).



(a) (b)

Figure 5.8: TYPE_VERB variable importance assessment for (a) stative nouns and (b) psychological states prediction.

Several key observations are as follows. First, comparable importance. The "mean decrease" in accuracy values for both TV1 TL and TV2 VA are relatively close across both 'Ph' (1.651 and 1.446, respectively) and 'Ps' (1.673 and 1.482) categories. This suggests that both true light verbs and vague action verbs play a somewhat similar role in predicting whether a noun within an LVC denotes a physiological behavior or a physiological state. Second, Slight Preference for TV1 TL in Predicting Physiological States. While the differences are marginal, TV1 TL exhibits a slightly higher mean decrease in accuracy for 'Ps' (1.673) compared to TV2 VA (1.482). This implies that true light verbs might offer a slightly stronger indication of a stative noun representing a physiological state, rather than a behavior. Third, overall Moderate Importance. The overall "mean decrease in accuracy" values for both verb types (1.658 for TV1 TL and 1.459 for TV2 VA) are relatively low compared to other potential predictors (not shown in this table). This suggests that while verb type does contribute to the classification of nouns into 'Ph' and 'Ps' categories, other linguistic factors play a more significant role.

As for the linguistic implications, concerning the verb semantics and noun classification, the relatively similar importance of TV1 TL and TV2 VA suggests that the semantic distinction between these verb types might not be the primary factor in determining whether a noun within an LVC denotes a physiological behavior or state. Other linguistic cues, such as the semantic properties of the noun itself or the syntactic context of the LVC, could be more influential in this classification. Also, the slightly higher importance of TV1 TL in predicting 'Ps' hints at a potential subtle tendency for true light verbs to collocate more frequently with nouns denoting physiological states. This could be attributed to the fact that true light verbs often contribute less semantic content to the LVC, allowing the noun to carry the primary meaning, which might be more conducive to expressing states rather than actions.

Table 5.7: Variable importance from a Random Forest Model for predicting status of the subdivision of *state-nouns* (Part A) and *physiological state nouns* (Part B).

PART A				PART B				
Variables	**(Ph)**	**(Ps)**	**Mean decrease accuracy**	**Variables**	**(Se)**	**(Ms)**	**(Es)**	**Mean decrease accuracy**
TV1 TL	1.651	1.673	**1.658**	TV1 TL	0.000	2.206	2.788	**2.565**
TV2 VA	1.446	1.482	1.459	TV2 VA	0.000	0.438	1.041	0.750

Furthermore, based on the Table 5.7 (Part B) presents the results of a variable importance analysis derived from a Random Forest model, focusing specifically on the role of the TYPE_VERB variable in predicting the status of stative nouns related to physiological states within Indonesian LVCs. This analysis aims to understand the extent to which the type of verb, classified as either True Light Verb (TV1 TL) or Vague Action Verb (TV2 VA), influences the categorization of nouns into three sub-domains: Sensations (Se), Mental states (Ms), and Emotional states (Es).

Based on the results of evaluation, at least, there are four key observations. First, TV1 TL's Predominance in Predicting Emotional States (Es): The most striking observation is the significantly higher mean decrease in accuracy for TV1 TL in the 'Es' category (2.788), compared to its impact on 'Se' (0.000) and 'Ms' (2.206). This strongly suggests that true light verbs are particularly important in predicting stative nouns that denote emotional states within Indonesian LVCs. Second, TV2 VA's Limited Role: In contrast, TV2 VA exhibits minimal impact across all three sub-domains, with the highest mean decrease in accuracy being only 1.041 for 'Es'. This implies that vague action verbs are less reliable predictors of the stative noun's sub-domain compared to true light verbs. Third, Overall Importance Skewed Towards TV1 TL: The "overall mean decrease" in accuracy values further emphasize the greater significance of TV1 TL (2.565) compared to TV2 VA (0.750) in predicting the sub-domain of stative nouns. Fourth, 'Sensations' (Se) as an exception. The zero values for both TV1 TL and TV2 VA in the 'Se' category indicate that the verb type has no discernible impact on predicting stative nouns related to sensations. This suggests that other linguistic factors, such as the semantic properties of the noun itself or the syntactic context of the LVC, might play a more dominant role in this particular sub-domain.

5.3.2 Features of eventive nouns within Indonesian LVCs

The subsequent discussion aims to quantitatively assess the influence of data-based variables on the selection of eventive nouns within Indonesian LVCs. These data-based variables encompass quantitative metrics such as the frequency of each LVC across the four examined corpora and their corresponding ranks, providing a measure of their empirical prominence. Furthermore, the classification of verbs into the categories of "true light verbs' (TVLs) and 'vague action verbs' (VAVs) introduces a crucial qualitative dimension to the analysis. This multifaceted approach seeks to address the fundamental question of whether the verbal element exerts a discernible influence on the selection of noun types within the context of Indonesian LVCs. By systematically examining the interplay between these quantitative and qualitative variables, we aim to uncover potential correlations and dependencies that shed light on the underlying linguistic mechanisms governing the formation and usage of LVCs.

Table 5.8: Variable importance from Random Forest model for predicting the status of *stative* and *eventive nouns* (mix of features).

Variables	Stative nouns	Eventive nouns	Mean decrease accuracy
Total Freq	5.894	11.303	17.234
Rank	3.602	12.967	**19.708**
TYPE VERB: TLVs | VAVs	9.393	15.569	17.581

In particular, Table 5.6 presents the results of a variable importance analysis originated from a Random Forest model, aiming to predict whether a noun within an Indonesian LVCs is stative (representing a state or condition) or eventive (representing an action or event). The model incorporates a mix of quantitative and qualitative variables, offering insights into the diverse linguistic factors influencing noun classification within LVCs. There are several key observations, specifically in relation to the eventive nouns. First, `TYPE_VERB` as the strongest predictor for eventive nouns. The `TYPE_VERB` variable exhibits the moderate mean decrease in accuracy (17.581) and highest connection to eventive nouns (15.569), indicating its substantial role in differentiating between stative and eventive nouns. This suggests that the semantic or syntactic class of the verb within the LVC is a crucial factor in determining the nature of the eventive noun.

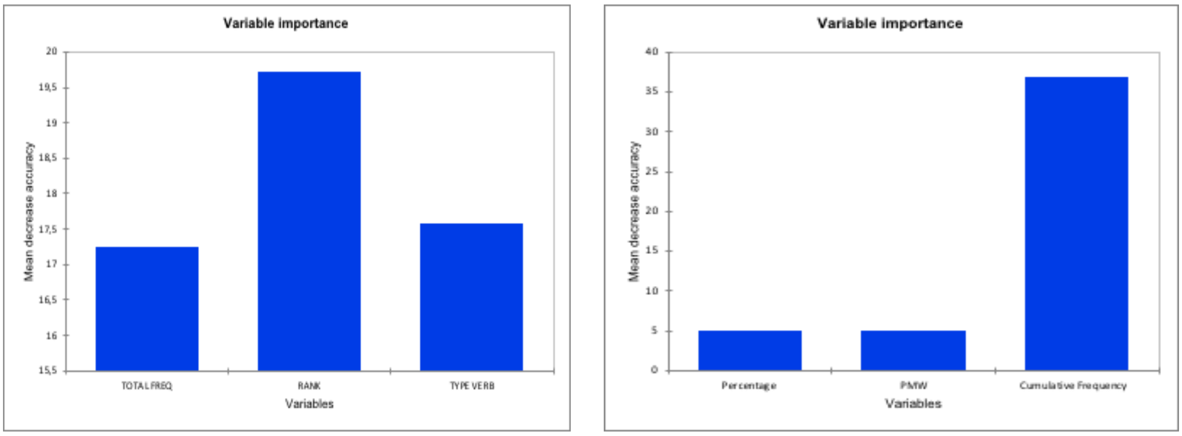
Second, quantitative variables' contribution. Both `TOTAL_FREQ` and `RANK` demonstrate moderate predictive power, with mean decrease accuracy values of 17.234 and 19.708, respectively. This implies that the frequency and relative ranking of an LVC in the corpus provide valuable information for noun classification. Also, regarding the frequency vs. rank, interestingly, `RANK` appears slightly more important than `TOTAL_FREQ` in predicting both stative and eventive nouns. This suggests that the relative frequency of an LVC, compared to other LVCs in the corpus, might be a more informative predictor than its absolute frequency. As of nuanced roles, while all variables contribute to the model's predictive accuracy, their specific impact on stative and eventive nouns varies. `TOTAL_FREQ` and `RANK` show a relatively balanced impact on both noun classes. `TYPE_VERB` seems to be more crucial in predicting eventive nouns (Ev: 15.569) than stative nouns (St: 9.393), suggesting that the verb's characteristics are particularly informative for identifying eventive nouns.

Table 5.9: Variable importance from Random Forest model for predicting the status of *stative* and *eventive nouns* (specific quantitative of features).

Variables	Stative nouns	Eventive nouns	Mean decrease accuracy
Percentage	4.684	0.758	4.891
PMW	4.550	-1.143	4.838
Cumulative Frequency	29.365	30.942	**36.879**

Furthermore, Table 5.7 presents the results of a variable importance analysis derived from a Random Forest model, aimed at predicting whether a noun within an Indonesian LVCs is stative or eventive. It specifically focuses on the predictive power of three quantitative variables: `Percentage`, `PMW`, and `Cumulative Frequency`. According to the evaluation, three notable key observations are as follows. First, `Cumulative Frequency` as the Dominant Predictor. `Cumulative Frequency` emerges as the most influential predictor, with the highest mean decrease in accuracy (36.879). Its substantial impact on both Stative nouns (29.365) and Eventive nouns (30.942) underscores its crucial role in differentiating between the two. This suggests that the overall distribution of LVC frequencies in the corpus provides valuable information for noun classification. It could imply that more frequent LVCs, regardless of their specific verb-noun combination, tend to exhibit certain patterns that aid in distinguishing stative from eventive nouns.

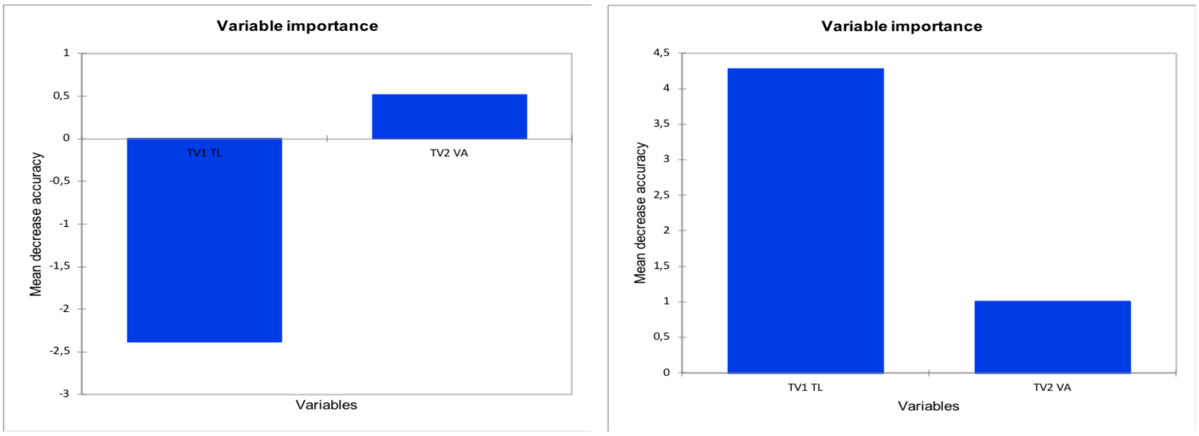
Second, `Percentage` and `PMW`. Both `Percentage` and `PMW` demonstrate moderate predictive power, with mean decrease accuracy values of 4.891 and 4.838, respectively. This implies that the relative frequency of an LVC and the strength of association between the verb and noun also contribute to noun classification, albeit to a lesser extent than cumulative frequency. Third, PMW and Eventive Nouns. Interestingly, PMW has a negative value (-1.143) for eventive nouns. This suggests that disrupting the PMW values actually improves the model's ability to identify eventive nouns. This counter-intuitive result might indicate that LVCs with strong verb-noun collocations (high PMW) are less likely to contain eventive nouns. It could imply that eventive nouns tend to occur in more diverse and less predictable combinations with verbs.



(a) (b)

Figure 5.9: Random Forest variable importance for *stative* and *eventive noun* prediction: (a) mixed feature set and (b) specific quantitative features.

In essence, regarding these specific quantitative variables in performance to predict the eventive noun within Indonesian LVCs, at least, there are three implications. As of frequency and distribution as key factors, the prominence of Cumulative Frequency suggests that the overall frequency distribution of LVCs in the corpus plays a significant role in noun classification. This could be attributed to the fact that more frequent LVCs are likely to be more conventionalized and exhibit clearer semantic or syntactic patterns that aid in distinguishing stative and eventive nouns. Also, of verb-noun collocation, the moderate importance of PMW indicates that the strength of the verb-noun association provides some predictive power, but it is not the most crucial factor. The negative impact of PMW on eventive noun prediction suggests a more nuanced relationship between collocational strength and noun classification, potentially reflecting the greater semantic flexibility of eventive nouns. Lastly, concerning relative frequency, the contribution of Percentage suggests that the relative frequency of an LVC within the corpus also plays a role in noun classification. This could indicate that more frequent LVCs, even if not highly ranked, may exhibit certain characteristics that aid in distinguishing stative and eventive nouns.



(a) (b)

Figure 5.10: `TYPE_VERB` variable importance assessment for (a) definite process nouns and (b) punctual nouns prediction.

Furthermore, upon delving deeper into the sub-classification of eventive nouns, we observe noteworthy patterns in the predictability of the `TYPE_VERB` for both definite process nouns and punctual nouns (Figure 5.11). Table 15 (Part A) presents the results of a variable importance analysis stemming from a Random Forest model. This analysis specifically examines the influence of the `TYPE_VERB` variable in predicting the class status of definite process nouns, a subcategory of eventive nouns representing specific and identifiable actions or events, within Indonesian LVCs. It delves into how the type of verb, categorized as either True Light Verb (`TV1 TL`) or Vague Action Verb (`TV2 VA`), impacts the classification of nouns into three distinct sub-domains: Cognitive Process (`Co`), Physical Activities (`Pa`), and Verbal Communication (`Vc`).

Based on the observations, there are distinctive notes as follows. First, `TV2 VA`'s predominance in predicting cognitive processes (`Co`). The most striking observation is the positive mean decrease in accuracy for `TV2 VA` in the '`Co`' category (1.427), contrasting with the negative value for `TV1 TL` (-0.060). This strongly indicates that vague action verbs are particularly crucial in predicting definite process nouns associated with cognitive processes within Indonesian LVCs. The negative value for `TV1 TL` suggests that its permutation might even slightly improve the model's performance, hinting at a potential inverse relationship between true light verbs and nouns denoting cognitive processes. Second, limited and mixed impact on physical activities (`Pa`) and Verbal Communication (`Vc`). Both `TV1 TL` and `TV1 VA` exhibit negative mean decrease in accuracy values for '`Pa`' and '`Vc`', with `TV1 TL` showing a more pronounced negative impact. This implies that permuting the verb type, particularly for true light verbs, might lead to a slight improvement in the model's ability to predict nouns related to physical activities and verbal communication. This observation suggests that other linguistic factors, beyond verb type, may play a more dominant role in classifying nouns within these sub-domains.

Third, Overall Importance Favors TV2 VA. The overall mean decrease in accuracy values underscores the greater significance of TV2 VA (0.510) compared to TV1 TL (-2.377) in predicting the sub-domain of definite process nouns. This further supports the notion that vague action verbs play a more crucial role in classifying these nouns, particularly those related to cognitive processes. Admittedly, some linguistic implications can be formulated. Regarding the Semantic Relevance of Vague Action Verbs. The strong association between TV2 VA and the 'Co' category suggests that vague action verbs in Indonesian might have a greater affinity for expressing definite process nouns related to cognitive processes. This could be attributed to their richer semantic content and their ability to convey the nuances of mental actions or events.

In respect to the complex role of True Light Verbs, the negative impact of TV1 TL permutation on predicting 'Pa' and 'Vc' nouns, and its minimal influence on the 'Co' category, hints at a complex relationship between true light verbs and the classification of definite process nouns. This suggests that true light verbs might be less informative in distinguishing between different sub-domains of definite process nouns, potentially due to their more grammatical and less semantically rich nature. Lastly, contextual and noun-specific factors. The mixed results for 'Pa' and 'Vc' categories emphasize the importance of considering additional linguistic features, such as the semantic properties of the noun itself and the syntactic context of the LVC, in classifying definite process nouns.

Moreover, regarding the predictability model evaluation for the punctual nouns, Table 5.10 (Part B) presents the results of a variable importance analysis derived from a Random Forest model. This analysis specifically focuses on the role of the TYPE_VERB variable in predicting the class status of punctual nouns within Indonesian LVCs. Punctual nouns, as previously established, capture momentary events or actions, and the table further categorizes these nouns into five sub-domains: Motion (Mo), Non-verbal Communication (Nc), Gesture (Ge), Vocal Sound (Vs), and Blow (Bl). The aim is to understand the extent to which the type of verb, classified as either True Light Verb (TV1 TL) or Vague Action Verb (TV2 VA), influences the categorization of nouns into these specific sub-domains.

Table 5.10: Variable importance from Random Forest model for predicting *definite-process* and *punctual noun* class status (Part A and B: TYPE_VERB features).

	PART A				**PART B**					
Variables	**Co**	**Pa**	**Vc**	**Mean decrease accuracy**	**Mo**	**Nc**	**Ge**	**Vs**	**Bl**	**Mean decrease accuracy**
TV1 TL	-0.060	-2.045	-1.348	-2.377	4.884	3.102	0.887	-0.170	0.000	**4.280**
TV2 VA	1.427	0.427	-1.418	**0.510**	0.000	1.005	0.000	1.005	0.000	1.005

For the notable observations, at least there are four notes. First, TV1 TL's Predominance in Predicting Motion (Mo) and Non-Verbal Communication (Nc). The most salient observation is the significantly higher mean decrease in accuracy for TV1 TL in the 'Mo' (4.884) and 'Nc' (3.102) categories compared to its impact on the other sub-domains. This strongly suggests that true light verbs are particularly crucial in predicting punctual nouns that denote motion and non-verbal communication within Indonesian LVCs. Second, TV2 VA's limited role, Except for Non-Verbal Communication (Nc) and Vocal Sound (Vs): In contrast, TV2 VA exhibits a negligible impact on most sub-domains, except for 'Nc' (1.005) and 'Vs' (1.005), where it shows a minor influence. This implies that vague action verbs are generally less reliable predictors of the specific sub-domain of punctual nouns, with some exceptions in the realms of non-verbal communication and vocal sounds.

Third, Overall Importance Favors TV1 TL. The overall mean decrease in accuracy values further emphasizes the greater significance of TV1 TL (4.280) compared to TV2 VA (1.005) in predicting the sub-domain of punctual nouns. Lastly, Gesture (Ge) and Blow (Bl) as Exceptions. The near-zero values for both TV1 TL and TV2 VA in the 'Ge'and 'Bl' categories indicate that verb type has minimal impact on predicting punctual nouns related to gestures and blows. This suggests that other linguistic factors, such as the semantic properties of the noun itself or the syntactic context of the LVC, might play a more dominant role in these particular sub-domains.

Accordingly, the linguistic implications are as follows. As of semantic and syntactic nuances, the strong association between TV1 TL and the 'Mo' and 'Nc' categories suggests that true light verbs in Indonesian may have a greater affinity for expressing punctual nouns related to motion and non-verbal communication. This could be attributed to their tendency to provide a grammatical framework while allowing the noun to carry the primary semantic load, which may be particularly suitable for expressing concrete actions and communicative gestures.

Additionally, on limited predictive power of vague action verbs, the relatively low impact of TV2 VA across most sub-domains indicates their lesser significance in predicting the specific category of punctual nouns. This could be due to their broader semantic range and their ability to combine with various types of nouns, potentially obscuring the distinction between different punctual noun categories. Lastly, contextual factors in 'Ge' and 'Bl’. The negligible impact of verb type on the ‘Ge’ and ‘Bl’ categories suggests that the classification of punctual nouns related to gestures and blows relies heavily on contextual cues or the specific semantic properties of the nouns themselves.

5.4 Resume

Chapter 5 presents a systemic investigation into the internal semantics of noun elements in Indonesian LVCs, framed through the dual taxonomy of stative and eventive nouns. Stative nouns are further categorized into psychological and physiological states. Eventive nouns are subdivided into indefinite, definite, and punctual process types. Through kernel density estimation, the chapter reveals frequency-based distinctions and divergence patterns across noun types. Importantly, machine learning models—particularly Random Forest—are employed to assess the predictive power of morphosyntactic and distributional features. Results show that eventiveness is strongly influenced by verb type and corpus distribution, while stative interpretations are more noun-driven. This chapter advances our understanding of how semantic weight, aspectual profile, and grammatical structure interact to shape LVC behavior in Indonesian.

Chapter 6

Conclusion

6.0 Outline

This chapter offers a detailed overview of the limitations encountered during and within the present study (§6.1). It is essential to acknowledge these limitations as the may have impacted the results and to provide readers with a clear understanding of the scope and implications of the study. Furthermore, the chapter concludes with some valuable recommendations for future research (§6.2). These recommendations are intended to guide future researchers who may wish to expand upon the study and explore new avenues for research. In short, this chapter is an indispensable part of the study, providing readers with an insightful stage for future research in this field.

6.1 Limitations of study

While this study offers insights into the intricate workings of Indonesian LVCs, it is essential to acknowledge its inherent limitations. These limitations, stemming from methodological and theoretical considerations, provide a critical lens through which to interpret the findings and suggest avenues for future research.

6.1.1 Data limitations

One of the inherent challenges in investigating LVCs in Indonesian using a corpus-based approach lies in corpus selection. The selection can significantly impact the generalizability and representativeness of the findings. A corpus restricted to a single register, such as formal written text, might overlook the rich variation of LVC usage evident in informal spoken language or creative writing. For instance, frequent LVCs found in business documents, like '*melakukan tindakan*' (take an action), might less prevalent in casual conversations where simples verbs like '*bertindak*' (to act) dominate.

Similarly, idiomatic LVCs specific to literary works might be absent from general news corpus. To achieve more comprehensive picture of LVCs distribution and frequency, the research should ideally leverage corpora encompassing a diverse range of registers. This could involve including sub-corpora from news articles, social media interactions, fictional narratives, conversational transcripts, and academic journals. By incorporating such a multifaceted corpus, the analysis can capture the nuanced ways LVCs function across different communicative contexts in Indonesian. This not only provides a more accurate reflection of how LVCs are employed in everyday life but also allows for a deeper understanding of how register variation influences their selection and semantic contribution. However, it is crucial to acknowledge the potential challenges associated with working with a large, multi-register corpus. These include increased time demands for corpus preparation, potential inconsistencies in annotation schemes across sub-corpora, and the need for more sophisticated corpus analysis tools to handle the inherent heterogeneity of the data. Despite these complexities, the benefits of a register-sensitive corpus design outweigh the drawbacks, ultimately leading to a more robust and insightful investigation of LVC phenomena in Indonesia.

Furthermore, a critical consideration is corpus size. While readily available smaller corpora may seem convenient, they can pose limitations in accurately capturing the full spectrum of LVC distribution and frequency. LVCs, by their nature, can exhibit a high degree of variability. Less common LVC types, particularly those specific to specialized domains or emerging trends, might be underrepresented or even entirely absent in a smaller corpus. This can lead to skewed results and an incomplete picture of LVC usage in Indonesian. On the other hand, to achieve statistically robust findings, the research should ideally leverage a large corpus. This allows for a greater chance of encountering a wider range of LVC types, including those with lower frequencies. For instance, a corpus focused on business communication might capture frequent LVCs like '*menjalankan fungsi*' (perform a function) but might miss rarer, idiomatic constructions like '*menggendong beban*' (carry a burden, meaning to take responsibility). A larger corpus, encompassing various registers, is more likely to include both common and less frequent LVCs, providing a more accurate reflection of their distribution across the language. However, simply increasing corpus size is not enough. For reliable statistical analysis, the corpus also needs to be diverse. A large corpus composed solely of formal written text, for example, would like to miss the prevalence of certain LVCs characteristic of informal speech. Therefore, the ideal corpus should be not only large but also encompass a variety of registers. This diversity ensures that the analysis captures LVC usage across different communicative contexts, leading to a more nuanced understanding of their frequency patterns.

In addition, a significant hurdle lies in the inherent subjectivity associated with annotation. Unlike full verbs with clear semantic content, LVCs often rely on a complex interplay between the light verb and its complement to convey meaning. This ambiguity can lead to challenges in consistently identifying and classifying LVCs within a corpus, particularly for borderline cases. For instance, the verb '*memberikan*' (give) can function as a full verb in some contexts as in (6.1) but can also be part of an LVC when paired with specific nouns as in (6.2), i.e. '*memberikan penjelasan*' (give an explanation). Annotators might disagree on whether such instances represent a full verb or an LVC, potentially introducing inconsistencies into the corpus data.

(6.1)	*Dia*	***memberikan***	***tiket***	*dengan*	*gemetaran*	*pada*	*kakaknya.*
	he/she-3.SG	give	ticket	with	shaking-ADJ	on-ADP	brother-BEN.POSS
	'He/she handed the ticket shakily to his brother.'						

(6.2)	*Aku*	*hanya*	*diam*	*saja*	*setelah*	***memberikan***	***penjelasan.***
	I-1.SG	only	silent	just	after	give	explanation
	'I just remined silent after giving the explanation.'						

To mitigate this subjectivity and enhance the reliability of the analysis, the research should establish clear and well-defined criteria for LVC identification. This could involve drawing upon existing theoretical frameworks for LVC analysis, adapting them to the specificities of Indonesian, and providing detailed operational definitions for annotators. Additionally, the criteria should address the semantic contribution of the light verb itself, the nature of the complement noun, and any potential syntactic markers that differentiate LVCs from full verb constructions.

6.1.2 Analysis-procedure limitations

While corpus analysis offers a powerful tool for investigating LVCs in Indonesian, the capabilities of the chosen analysis tools can significantly impact the types of analysis that can be performed. Basic concordance programs, for instance, while valuable for initial exploration, might not be sufficient for delving deeper into the intricacies of LVC structure and function. To achieve a more comprehensive understanding, the research should consider utilizing advance corpus analysis tools designed to handle complex linguistic features like LVCs. These advanced tools often offer functionalities that go beyond simple word searches. Features like part-of-speech tagging and dependency parsing can be instrumental in identifying the light verb and its complement within an LVC, allowing for a more precise analysis of their syntactic relationships.

Additionally, tools with pattern matching capabilities can be invaluable for uncovering specific LVC patterns or co-occurrence frequency with particular complement nouns. This can shed light on preferred verb-noun combinations and the semantic nuances associated with different LVC types. However, it is important to acknowledge that even advanced corpus analysis tools have limitations. While they can automate certain tasks and provide valuable insights, they might not be able to fully capture the subtleties of human language processing. For instance, identifying idiomatic LVCs or those with metaphorical interpretations might still require manual review and analysis by the researcher. Additionally, some tools might struggle with the inherent ambiguity associated with LVCs, particularly in borderline cases. This highlights the importance of critical evaluation and potentially employing complementary methodologies alongside corpus analysis tools.

Moreover, while corpus analysis offers a wealth of quantitative data on the frequency of LVCs in Indonesian, relying solely on frequency counts can provide an incomplete picture. Frequency data, undoubtedly valuable, reveals patterns in LVC usage but does not necessarily capture the full range of factors influencing their selection and function in different communicative contexts. To achieve a more nuanced understanding, the research should consider employing qualitative analysis alongside quantitative data from the corpus. Qualitative analysis allows for a deeper exploration of the contextual and functional nuances associated with LVC usage. This can involve manually examining specific examples of LVCs within the corpus, paying close attention to surrounding elements like register, discourse structure, and thematic focus. By delving into these contextual details, the research can uncover how factors like formality, genre, and speaker intention influence the choice of LVC over a full verb. For instance, the LVC '*melakukan tindakan*' (take action) might be more frequent in formal reports, whereas a more informal conversation might favor the simpler verb '*bertindak*' (to act). By integrating quantitative data on frequency with qualitative analysis of context and function, the research can develop a more holistic understanding of LVC usage in Indonesian. This combined approach can reveal not only how often LVCs appear but also why they are chosen and how they contribute to the overall meaning and communicative intent within different discourse contexts.

6.1.3 Theoretical-framework limitations

While this study is firmly anchored in established linguistic frameworks, it is crucial to acknowledge that the theoretical landscape of LVCs research is dynamic and evolving. The interpretations and conclusions presented herein should be considered within the context of this ongoing theoretical discourse, recognizing the potential for future refinements and alternative perspectives. Indeed, the very nature of LVCs, with their intricate interplay between lexical semantics and syntax, invites exploration from diverse angles.

One avenue for future research lies in exploring alternative theoretical frameworks that might offer complementary or contrasting perspective on Indonesian LVCs. For instance, Cognitive Linguistics, with its emphasis on embodied cognition and conceptual metaphor, could provide valuable insights into the conceptual motivations underlying the formation and interpretation of LVCs. This perspective might shed light on the cognitive processes involved in mapping abstract notions onto concrete verbs and how these mapping contribute to the expressive function of LVCs. Similarly, Construction Grammar, with its focus on the form-meaning pairings of linguistic unites, could offer a nuanced understanding of the specific constructions and their associated meanings within the domain of Indonesian LVCs. This approach might reveal subtle distinctions between seemingly similar LVCs, highlighting the role of construction-specific constraints and affordances in shaping their interpretation and usage.

6.2 Recommendations

6.2.1 Expanding the scope of LVCs analysis

The present study on LVCs in contemporary Indonesian offers a valuable snapshot of their usage patterns. However, a more comprehensive understanding of the language’s development and evolving communication patterns can be achieved by conducting a diachronic analysis that delves into the historical evolution of LVCs. Such an approach can reveal insights into how these constructions have emerged, solidified, or even undergone grammaticalization processes within Indonesian over time. Examining LVCs usage across different historical periods can also provide a deeper understanding of their semantic nuances and grammatical functions. For instance, tracking the development of common LVCs like '*melakukan*' (lit. to do) or '*memberikan*' (lit. to give) can shed light on they have gained specific meanings or evolved their grammatical structures over time. By identifying the influences of other languages on LVC formation in Indonesian's historical development, we can also gain a better understanding of the language's cultural and linguistic evolution.

On the other hand, a diachronic approach can help identify historical shifts in LVC frequency and distribution, offering insights into broader trends in Indonesian grammar. For example, a decline in the use of certain LVCs might reflect a growing preference for simpler verb forms, providing clues about the language's increasing efficiency or analytical nature. Such insights can contribute to a more comprehensive picture of Indonesian's historical trajectory and its ongoing evolution as a dynamic communication system. Thus, a diachronic analysis of LVCs in Indonesian can provide a more nuanced understanding of their development, usage patterns, and evolution over time. By incorporating a historical perspective, future research can contribute to a more comprehensive and detailed understanding of the language and its cultural and linguistic evolution.

Moreover, Indonesian, despite its status as a national language, exhibits rich dialectal variation across its vats geographical spread. While this study has focused on LVCs within a standardized Indonesian framework, future research can significantly benefit from exploring these constructions across different dialect regions. Investigating LVC usage in dialects like Javanese Indonesian, Sumatran Indonesian, or Eastern Indonesian varieties can reveal fascinating insights into potential variations and regional specificities. This could involve identifying unique LVC type differ regionally or uncovering dialect-specific semantic nuances associated with particular light verbs. In this respect, examining dialectal variation in LVCs can shed light on the historical development and diversification of Indonesian itself. Certain LVCs might be remnants of older language forms preserved within specific dialects, offering valuable clues about the language's evolution across different geographical areas. Additionally, exploring dialectal variation can contribute to a more comprehensive understanding of how LVCs function within diverse communicative contexts, shaped by regional cultural practices and social interactions.

A deeper understanding can also be gained through multilingual comparison. Investigating LVC usage in Indonesian alongside other Southeast Asian languages, or languages with similar verb-argument structures, can reveal interesting typological similarities and differences. This comparative approach can shed light on the broader linguistic landscape of the region and illuminate potential historical connections or independent developments. For instance, comparing LVC usage in Indonesian with language like Malay or Javanese, which share historical roots, could reveal shared patterns of LVC formation or identify divergences that highlight the unique grammatical trajectory of Indonesian.

6.2.2 Deepening theoretical exploration

By incorporating insights from Cognitive Linguistics, future research on LVCs in Indonesian can delve deeper into the fascinating realm of mental processes involved in their formation and interpretation. Cognitive Linguistics emphasizes the link between language, thought, and the embodied experience of the world. In the context of LVCs, this framework can shed light on how the interplay between the light verb's inherent meaning, the complement noun, and the overall context shapes understanding within the human mind. One key area of exploration lies in examining how the semantic contribution of the light verb itself interacts with the conceptual meaning of the complement noun. While some light verbs might have relatively bleached semantic content, others can introduce subtle nuances of aspect, modality, or even metaphorical interpretations. Cognitive Linguistics can help us understand how these light verb meanings combine with the specific concept evoked by the complement noun to create a richer mental representation. For instance, the LVC '*memberikan penjelasan*' (lit. to give an explanation) goes beyond the simple act of giving. The light verb '*memberikan*' (lit. to give) contributes a sense of completion, while the complement '*penjelasan*' (lit. explanation) activates a specific mental space related to conveying information. By integrating cognitive linguistic frameworks, we can explore how these elements interact in the mind to construct a nuanced understanding of the LVC's meaning.

Furthermore, Cognitive Linguistic can help us understand how the broader context influences LVC interpretation. The surrounding discourse, shared knowledge between speaker and listener, and the overall communicative situation all play a role in shaping meaning. For example, the LVC '*melakukan tindakan*' (lit. to take action) might evoke a more forceful image in a military context compared to a more procedural interpretation in a business meeting.

Cognitive Linguistics allows us to explore how mental spaces activated by the context interact with the LVC's inherent meaning and the complement noun. By adopting this cognitive approach, future research can move beyond purely syntactic or semantic analyses of LVCs. It can delve into the fascinating realm of human cognition, exploring how LVCs reflect the way we categorize the world, construct mental spaces, and ultimately make sense of the information conveyed through language. This deeper understanding can not only enrich our knowledge of LVCs in Indonesian but also contribute to a broader understanding of the intricate relationship between language and thought.

Future research can also benefit from considering LVCs through the lens of Construction Grammar. This framework views language as a network of conventionalized form-meaning pairings, and it offers valuable tools for identifying and characterizing LVCs as specific units within Indonesian grammar. This framework emphasizes the importance of identifying constructions as wholes, rather than simply analyzing their component parts. In the context of LVCs, this approach allows us to move beyond the individual light verb and its complement, and instead focus on the LVC as a unified construction with its own characteristic form and meaning. By employing construction grammar frameworks, future research can delve deeper into identifying the specific formational patterns associated with LVCs in Indonesian. This could involve analyzing the syntactic slots available for complements (e.g., subject vs. object position) and the potential restrictions on the types of nouns that can be used with particular light verbs. Additionally, constructions grammar can help us identify any specific morphological markers or idiomatic expressions that further characterize LVCs as distinct constructions within the language.

Furthermore, Construction Grammar offers valuable tools for exploring the semantic and pragmatic functions of LVCs as constructions. This framework allows us to go beyond the sum of the individual verb and noun meanings and consider the specific meaning that emerges from the LVC construction itself. For instance, the LVC *memberikan kesan* (lit. to give an impression) conveys a metaphorical meaning of creating a certain image or feeling. Construction grammar allows us to analyze this metaphorical interpretation not just as a result of the light verb's contribution but as property of the entire LVC construction. Additionally, this framework can help us explore the pragmatic functions associated with LVC usage. Do certain LVC constructions convey formality, politeness, or emphasis? Construction grammar can guide us in identifying these pragmatic nuances as inherent features of specific LVC constructions in Indonesian. By adopting this framework, future research can move beyond a purely compositional analysis of LVCs. It can explore these constructions as holistic units with their own unique form-meaning pairngs and functional roles within Indonesian grammar. This approach can contribute to a richer understanding of how LVCs contribute to the overall structure, meaning-making, and communicative effectiveness of the language.

6.2.3 Refining corpus-analysis techniques

While corpus analysis has proven its value in this study of LVCs in Indonesian, the potential for further exploration lies in utilizing even more advanced corpus analysis tools. These tools can move beyond basic keyword searches and offer functionalities specifically designed for handling complex linguistic features like LVCs. By incorporating these advanced functionalities, future research can unlock new insights into LVC usage and distribution within the language. One particularly valuable feature to explore lies in advanced pattern matching capabilities. These tools can go beyond simple word searches and identify specific LVC patterns within the corpus. For instance, the ability to search for patterns like “Verb + Preposition + Noun” or identify specific light verbs followed by a wider range of complement noun types can reveal nuanced co-occurrence patterns and shed light on preferred verb-noun combinations within LVCs. Additionally, functionalities for identifying idiomatic or metaphorical LVC usage can be highly beneficial. These tools might employ advanced algorithms to recognize fixed expressions or patterns associated with idiomatic LVCs, allowing researchers to delve deeper into the non-literal meanings conveyed through these constructions.

Furthermore, advanced corpus analysis tools can offer functionalities for exploring the syntactic and semantic environment surrounding LVCs. Features like dependency parsing can reveal the grammatical relationships between the light verb, complement noun, and other elements within the sentence. This can provide valuable insights into how LVCs function within broader syntactic structures. Additionally, tools with semantic role labeling capabilities can identify the semantic roles played by the light verb and complement noun within the LVCs, offering a deeper understanding of their contribution to the overall meaning of the construction. By actively seeking out and utilizing these advanced functionalities within corpus analysis tools, feature research on LVCs in Indonesian can move beyond basic frequency analysis. It can delve into the intricacies of LVC structure, co-occurrence patterns, idiomatic usage, and the interplay between syntax and semantics within these fascinating constructions. Ultimately, this advanced approach can lead to a richer and more nuanced understanding of how LVCs function and contribute to expressive power of the Indonesian language.

Furthermore, enriching this study approach with methods from Discourse Analysis can offer a more comprehensive picture of how LVCs function within real-world communication. Discourse Analysis focuses on language use beyond the sentence level, considering the broader context and how utterances interact to create meaning. Integrating this perspective into future LVC research can illuminate how these constructions are shaped by and contribute to the flow of conversation. One key area of exploration lies in examining LVC usage within a larger conversational context. Corpus analysis often presents language in isolation, but real communication involves back-and-forth exchanges. By examining LVCs within dialogue or longer stretches of discourse, we can see how the choice of a particular LVC might be influenced by preceding utterances or the speaker's desire to connect to previous ideas. For instance, the LVC '*memberikan solusi*' (lit. to give a solution) might be chosen to directly respond to a problem earlier in the conversation. Discourse Analysis allows us to trace these connections and understand how LVCs contribute to the overall coherence and progression of a discourse.

Discourse Analysis can also help us consider the role of shared knowledge between speaker and listener in interpreting LVCs. The specific meaning conveyed by an LVC might not solely reside within the construction itself but can be enriched by the participant' mutual understanding of the situation or topic at hand. For example, the LVC '*melakukan penelitian'* (to conduct research) used in a scientific discussion carries different implications compared to its use in a casual conversation about a personal hobby. Discourse Analysis allows us to unpack these layers of meaning in use that emerge through shared knowledge and the context of the discourse. By incorporating this framework alongside corpus analysis, future research on LVCs can move beyond a purely quantitative or structural approach. It can delve into the dynamic nature of language use, exploring how LVCs function within the intricate web of conversation, shaped by both the linguistic context and the shared understanding between interlocutors. Specifically, future research should use genre-balanced corpora (e.g., carefully compiled Indonesian GNC-type corpora) to investigate how LVC patterns vary across discourse type. This multifaceted perspective can ultimately lead to a richer and more nuanced understanding of how LVCs contribute to meaning-making in everyday Indonesian communication.

References

Abeillé, A., & Godard, D. (2001). A Class of "Lite" Adverbs in French. In *Romance Syntax, Semantics and L2 Acquisition: Selected papers from the 30th Linguistic Symposium on Romance Languages*. John Benjamins Publishing Company. https://doi.org/10.1075/cilt.216.04abe

Acedo-Matellán, V., & Pineda, A. (2019). Light Verb Constructions in Basque and Romance. In *Grammars and Sketches of the World's Languages* (Vol. 7). https://doi.org/10.1163/9789004395398_007

Agee, M. A. (2018). The Collaborative Function of Verbal Aspect and Aktionsart: A Distributional Analysis of English Verb-Types. *Linguistics Senior Research Projects. 11*. https://digitalcommons.cedarville.edu/linguistics_senior_projects/11

Aikhenvald, A. Y. (2006). Classifiers and Noun Classes: Semantics. In *Encyclopedia of Language & Linguistics* (pp. 463–471). Elsevier. https://doi.org/10.1016/B0-08-044854-2/01111-1

Aikhenvald, A. Y. (2007). Typological distinctions in word-formation. In *Language Typology and Syntactic Description* (pp. 1–65). Cambridge University Press. https://doi.org/10.1017/CBO9780511618437.001

Aikhenvald, A. Y. (2017). A Typology of Noun Categorization Devices. In *The Cambridge Handbook of Linguistic Typology* (pp. 361–404). Cambridge University Press. https://doi.org/10.1017/9781316135716.012

Alexiadou, A., Anagnostopoulou, E., & Schäfer, F. (2015). External Arguments in Transitivity Alternations. *External Arguments in Transitivity Alternations*. https://doi.org/10.1093/ACPROF:OSO/9780199571949.001.0001

Alkoffash, M. S. (2012). Comparing between Arabic Text Clustering using K Means and K Mediods. *International Journal of Computer Applications, 51*(2), 5–8.

Altmann, G. (1966). Binomial index of euphony for Indonesian poetry. *Asian and African Studies, 2*, 62–67.

Altmann, G. (1967). The structure of Indonesian morphemes. *Asian and African Studies, 3*, 23–36.

Altmann, G. (1985). Sprachtheorie und mathematische Modelle. *SAIS Arbeitsberichte Aus Dem Seminar Für Allgemeine Und Indogermanische Sprachwissenchaft* 8, 1–13.

Altmann, G. (1997). The Art of Quantitative Linguistics. *Journal of Quantitative Linguistics*, 4(1–3), 13–22. https://doi.org/10.1080/09296179708590074

Altmann, G. (2025). Language theory and mathematical models. In E. Kelih, J. Mačutek, & M. Koščová (Eds.), *Quantification in Linguistics and Text Analysis: Selected Papers of Gabriel Altmann*. De Gruyter Mouton.

Altmann, E. G., & Gerlach, M. (2016). *Statistical Laws in Linguistics* (pp. 7–26). https://doi.org/10.1007/978-3-319-24403-7_2

Almodaresi, F., Ungar, L., Kulkarni, V., Zakeri, M., Giorgi, S., & Schwartz, H. A. (2017). On the Distribution of Lexical Features at Multiple Levels of Analysis. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 79–84. https://doi.org/10.18653/v1/P17-2013

Al-Azzawy, D. S., & Al-Rufaye, F. M. L. (2017). Arabic words clustering by using K-means algorithm. *Annual Conference on New Trends in Information & Communications Technology Applications (NTICT)*, 263–267. https://doi.org/10.1109/NTICT.2017.7976098

Andreou, M. (2017). The semantics of compounding. *Morphology*, 27(4), 721–725. https://doi.org/10.1007/s11525-017-9311-1

Angheluta, R., Jeuniaux, P., Mitra, R., & Moens, M.-F. (2004). Clustering Algorithms for Noun Phrase Coreference Resolution. *Journ´ees Internationales d'Analyse Statistique Des Donn´ees Textuelles, 7*, 60–69. https://www.i6doc.com/resources/titles/28001100205490

Arad, M. (2005). Roots: Where Syntax, Morphology, and the Lexicon Meet. In M. Arad (Ed.), *Roots and Patterns: Hebrew Morpho-syntax* (pp. 1–23). Springer Netherlands. https://doi.org/10.1007/1-4020-3244-7_1

Aradilla, G., Bourlard, H., & Doss, M. M. (2008). Using KL-based acoustic models in a large vocabulary recognition task. *Interspeech 2008*, 928–931. https://doi.org/10.21437/Interspeech.2008-110

Arora, A., Meister, C., & Cotterell, R. (2022). Estimating the Entropy of Linguistic Distributions. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 175–195. https://doi.org/10.18653/v1/2022.acl-short.20

Ausensi, J. (2021). The semantics of roots determines argument structure. *Proceedings of Sinn Und Bedeutung*, *25*(0), 95–111. https://doi.org/10.18148/sub/2021.v25i0.926

Baayen, H. (1989). *A Corpus-Based Approach to Morphological Productivity. Statistical Analysis and Psycholinguistic Interpretation*. Vrije Universiteit.

Baayen, H. (1992). *Quantitative aspects of morphological productivity* (pp. 109–149). https://doi.org/10.1007/978-94-011-2516-1_8

Baayen, H. (2009). Corpus linguistics in morphology: Morphological productivity. In *Corpus Linguistics* (pp. 899–919). Mouton de Gruyter. https://doi.org/10.1515/9783110213881.2.899

Baayen, H., & Lieber, R. (1991). Productivity and English derivation: a corpus-based study. *Ling*, *29*(5), 801–844. https://doi.org/10.1515/ling.1991.29.5.801

Bach, E. (1965). Structural Linguistics and the Philosophy of Science. *Diogenes*, *13*(51), 111–128. https://doi.org/10.1177/039219216501305107

Bach, E. (1988). Categorial Grammars as Theories of Language. In *Categorial Grammars and Natural Language Structures. Studies in Linguistics and Philosophy, vol 32* (pp. 17–34). Springer. https://doi.org/10.1007/978-94-015-6878-4_2

Bach, E. (2005). Eventualities, Grammar, and Language Diversity. In H. J., Verkuyl, H., de Swart, & A. van Hout (Eds.), *Perspectives on Aspect. Studies in Theoretical Psycholinguistics* (pp. 167–180). Springer. https://doi.org/10.1007/1-4020-3232-3_9

Bache, C. (1982). Aspect and Aktionsart: towards a semantic distinction. *Journal of Linguistics*, *18*(1), 57–72. https://doi.org/10.1017/S0022226700007234

Baerman, M., Brown, D., & Corbett, G. G. (2005). *The Syntax-Morphology Interface*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486234

Baker, P., & Egbert, J. (2016). Triangulating methodological approaches in corpus linguistic research. In *Triangulating Methodological Approaches in Corpus Linguistic Research*. https://doi.org/10.4324/9781315724812

Baker, P., Hardi, A., & McEnery, T. (2006). *Glossary of Corpus Linguistics*. Edinburgh University Press.

Barking, M., Backus, A., & Mos, M. (2022). Similarity in Language Transfer - Investigating Transfer of Light Verb Constructions From Dutch to German. *Journal of Language Contact*, *15*(1), 198–239. https://doi.org/10.1163/19552629-15010005

Baroni, M., & Evert, S. (2009). Statistical methods for corpus exploitation. In *Corpus Linguistics: An International Handbook* (Vol. 2).

Bauer, L., Lieber, R., & Plag, I. (2013). *The Oxford Reference Guide to English Morphology*. Oxford University PressOxford. https://doi.org/10.1093/acprof:oso/9780198747062.001.0001

Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i-Cancho, R. (2017). The Entropy of Words—Learnability and Expressivity across More than 1000 Languages. *Entropy*, *19*(6), 275. https://doi.org/10.3390/e19060275

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, *5*(4), 257–269. https://doi.org/10.1093/llc/5.4.257

Biber, D. (2012). Corpus-Based and Corpus-driven Analyses of Language Variation and Use. In *The Oxford Handbook of Linguistic Analysis*. https://doi.org/10.1093/oxfordhb/9780199544004.013.0008

Biber, D., & Jones, J. K. (2009). Quantitative methods in corpus linguistics. In *Corpus Linguistics: An International Handbook* (Vol. 2).

Binnick, R. I. (1991). *Time and the Verb*. Oxford University Press. https://doi.org/10.1093/oso/9780195062069.001.0001

Blevins, J. P. (2016). Word and Paradigm Morphology. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199593545.001.0001

Blom, C. (2005). *The Demarcation of Morphology and Syntax* (pp. 53–66). https://doi.org/10.1075/cilt.264.04blo

Bowman, A. W., & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford University Press. https://doi.org/10.1093/oso/9780198523963.001.0001

Booij, G. (2016). Construction Morphology. In *The Cambridge Handbook of Morphology* (pp. 424–448). Cambridge University Press. https://doi.org/10.1017/9781139814720.016

Bouveret, M. (2021). Lexicalization, grammaticalization and constructionalization of the verb give across languages A cognitive case study of language innovation. In *Constructional Approaches to Language* (Vol. 29). https://doi.org/10.1075/cal.29.int

Brugman, C. (2001). Light verbs and polysemy. *Language Sciences*, *23*(4–5), 551–578. https://doi.org/10.1016/S0388-0001(00)00036-X

Butt, M. (2010). The light verb jungle: Still hacking away. In *Complex Predicates: Cross-Linguistic Perspectives on Event Structure*. https://doi.org/10.1017/CBO9780511712234.004

Bygi, Z. R., Karimi-Doustan, G.-H., & Sharif, B. (2018). Explanation for alternating Light Verbs (LVs) in Persian complex predicates from a generative lexicon viewpoint. *Language Related Research*, 8(7), 429–452.

Plaza, A. B. (2005). Poner en movimiento/in Bewegung setzen:¿ verbos pseudocopulativos españoles frente a verbos funcionales alemanes?. In *Fraseología contrastiva: con ejemplos tomados del alemán, español, frances e italiano* (pp. 185-196). Servicio de Publicaciones de la Universidad de Murcia.

Caballero, G., & Inkelas, S. (2013). Word construction: tracing an optimal path through the lexicon. *Morphology*, *23*(2), 103–143. https://doi.org/10.1007/s11525-013-9220-x

Cai, H., Qu, Y., & Feng, Z. (2019). A Corpus-Based Study of the Semantic Prosody of Chinese Light Verb Pattern across Registers: Taking jinxing and shoudao as Examples. *Glottometrics, 46*, 61-82.

Caro, E. M., & Arús-Hita, J. (2020). Give as a light verb. *Functions of Language*, *27*(3), 280–306. https://doi.org/10.1075/fol.16036.mar

Chen, B. (2018). The Influence of Hermann Paul's Linguistic Ideas after the First Publication of Principien der Sprachgeschichte (1880). *Amsterdamer Beiträge Zur Älteren Germanistik*, *78*(2–3), 313–335. https://doi.org/10.1163/18756719-12340121

Chen, Y.-C. (2017). A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology, 1*(1), 161–187. https://doi.org/10.1080/24709360.2017.1396742

Cieri, C., Maxwell, M., Strassel, S. and Tracey, J. (2016). Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549.

Coleman, L., & Kay, P. (1981). Prototype Semantics: The English Word Lie. *Language, 57*(1), 26. https://doi.org/10.2307/414285

Comrie, B. (1976). The Syntax of Causative Constructions: Cross-Language Similarities and Divergences. In *the Grammar of Causative Constructions* (pp. 259–312). BRILL. https://doi.org/10.1163/9789004368842_011

Comrie, B. (1993). Language universal and linguistic typology: data-bases and explanations. *STUF - Language Typology and Universals*, *46*(1–4). https://doi.org/10.1524/stuf.1993.46.14.3

Comrie, B. (1997). On the origin of the Basic Variety. *Second Language Research*, *13*(4), 367–373. https://doi.org/10.1191/026765897668475347

Comrie, B. (2000). From potential to realization: an episode in the origin of language. *Linguistics*, *38*(5). https://doi.org/10.1515/ling.2000.019

Comrie, B. (2017). Languages of the World. In *The Handbook of Linguistics* (pp. 21–38). Wiley. https://doi.org/10.1002/9781119072256.ch2

Considine, J. (2014). Samuel Johnson and Johann Christoph Adelung. In *Academy Dictionaries 1600–1800* (pp. 121–143). Cambridge University Press. https://doi.org/10.1017/CBO9781107741997.008

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press.

Dai, G. (2016). Light Verbs in Translated Chinese. In *Hybridity in Translated Chinese. New Frontiers in Translation Studies* (pp. 133–154). Springer. https://doi.org/10.1007/978-981-10-0742-2_9

Dal, G., & Namer, F. (2015). Frequency in morphology: For what usages? | La fréquence en morphologie: Pour quels usages? *Langages, 197*(1), 47–68. https://doi.org/10.3917/lang.197.0047

Di, J., & Gou, X. (2018). Bisecting K-means Algorithm Based on K-valued Selfdetermining and Clustering Center Optimization. *Journal of Computers,* 588–595. https://doi.org/10.17706/jcp.13.6.588-595

Diessel, H. (2017). Usage-Based Linguistics. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.363

Divjak, D., & Fieller, N. (2014). Cluster analysis. In D. Glynn & J. A. Robinson (Eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy* (pp. 405–441). John Benjamins Publishing Company. https://doi.org/10.1075/hcp.43.16div

Eatough, V., & Tomkins, L. (2022). Qualitative methods. In *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science [HSK] 46*/1 (pp. 163–182). De Gruyter. https://doi.org/10.1515/9783110347524-008

Egbert, J., Larsson, T., & Biber, D. (2020). Doing linguistics with a corpus: Methodological considerations for the everyday user. In *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User*. https://doi.org/10.1017/9781108888790

Egghe, L. (2007). Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments. *Journal of the American Society for Information Science and Technology*, *58*(5), 702–709. https://doi.org/10.1002/asi.20524

Egghe, L., & Rousseau, R. (2003). Size-frequency and rank-frequency relations, power laws and exponentials: a unified approach. *Progress in Natural Science*, *13*(6), 478–480. https://doi.org/10.1080/10020070312331343880

Embick, D. (2013). Morphemes and morphophonological loci. *Distributed Morphology Today: Morphemes for Morris Halle*, 151–166. https://doi.org/10.7551/MITPRESS/9780262019675.003.0009

Embick, D. (2015). *The Morpheme*. De Gruyter. https://doi.org/10.1515/9781501502569

Embick, D., & Noyer, R. (2007). *Distributed Morphology and the Syntax—Morphology Interface*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199247455.013.0010

Eshaghi, M., & Karimi-Doostan, G. (2023). Persian Light Verbs as event determiners. In *Light Verb Constructions as Complex Verbs* (pp. 73–98). De Gruyter. https://doi.org/10.1515/9783110747997-004

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis.* Wiley. https://doi.org/10.1002/9780470977811

Falk, J. S. (1992). Otto Jespersen, Leonard Bloomfield, and American Structural Linguistics. *Language*, *68*(3), 465. https://doi.org/10.2307/415791

Fleischhauer, J. (2021). Light Verb Constructions and Their Families – A Corpus Study on German stehen unter-LVCs. *MWE 2021 - 17th Workshop on Multiword Expressions, Proceedings of the Workshop*, 63–69.

Fleischhauer, J. (2023). Prospective Aspect and Current Relevance: A Case Study of the German Prospective Stehen vor NP Light Verb Construction. *Journal of Germanic Linguistics*, *35*(4), 371–408. https://doi.org/10.1017/S1470542722000198

Fleischhauer, J., & Gamerschlag, T. (2014). We're going through changes: How change of state verbs and arguments combine in scale composition. *Lingua*, *141*, 30–47. https://doi.org/10.1016/j.lingua.2013.01.006

Fleischhauer, J., & Neisani, M. (2020). Adverbial and attributive modification of Persian separable light verb constructions. *Journal of Linguistics*, *56*(1), 45–85. https://doi.org/10.1017/S0022226718000646

Fleischhauer, J., Gamerschlag, T., Kallmeyer, L., & Petitjean, S. (2019). Towards a compositional analysis of German light verb constructions (LVCs) combining Lexicalized Tree Adjoining Grammar (LTAG) with frame semantics. *IWCS 2019 - Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, 79–90.

Frawley, W. J. (2003). Bopp, Franz. In *International Encyclopedia of Linguistics$ International Encyclopedia of Linguistics* (2nd ed.). Oxford University Press.

Fries, C. C. (1954). Meaning and Linguistic Analysis. *Language*, *30*(1), 57. https://doi.org/10.2307/410220

Fries, C. C. (1961). Advances in Linguistics. *College English*, *23*(1), 30. https://doi.org/10.2307/373935

Fries, P. H. (2008). Charles C. Fries, Linguistics And Corpus Linguistics. *The 19th European Systemic Functional Linguistics Conference and Workshop.*

Gale, W. A., & Sampson, G. (1995). Good-turing frequency estimation without tears*. *Journal of Quantitative Linguistics*, *2*(3), 217–237. https://doi.org/10.1080/09296179508590051

García-Pardo, A. (2021). Light Verbs and the Syntactic Configurations of se. In *Studies in Natural Language and Linguistic Theory* (Vol. 99). https://doi.org/10.1007/978-3-030-57004-0_10

Giparaitė, J. (2024). A corpus-based analysis of light verb constructions with MAKE and DO in British English. *Kalbotyra, 76,* 18–41. https://doi.org/10.15388/Kalbotyra.2023.76.2

Giparaitė, J. (2015). A Corpus-based Analysis of the Constructions Have/Take/Get a Bath and Have/Take/Get a Rest in British English. *Žmogus Ir Žodis, 17*(3), 37–53. https://doi.org/10.15823/zz.2015.10

Goddard, C. (2001). Lexico-Semantic Universals: A Critical Overview. *Linguistic Typology, 5*(1), 1–65. https://doi.org/10.1515/lity.5.1.1

Goddard, C., Wierzbicka, A. (2016) *Words and Meanings. Lexical Semantics Across Domains, Languages and Cultures*. Oxford University Press, London.

Goldhahn, D., Eckart, T. & Quasthoff, U. (2012) Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In: *Proceedings of the 8th International Language Resources and Evaluation (LREC'12).*

Good, I. J. (1953). The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika*, *40*(3–4), 237–264. https://doi.org/10.1093/biomet/40.3-4.237

Gramacki, A. (2018). *Nonparametric Kernel Density Estimation and Its Computational Aspects* (Vol. 37). Springer International Publishing. https://doi.org/10.1007/978-3-319-71688-6

Greenberg, J. H. (1956). The Measurement of Linguistic Diversity. *Language*, *32*(1), 109. https://doi.org/10.2307/410659

Gries, S. T. (2015a). Quantitative designs and statistical techniques. In *the Cambridge Handbook of English Corpus Linguistics*. https://doi.org/10.1007/9781139764377.004

Gries, S. T. (2015b). Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language and Linguistics, 16*(1), 93–117. https://doi.org/10.1177/1606822X14556606

Gries, S. Th., & Paquot, M. (2020). Writing up a Corpus-Linguistic Paper. In *A Practical Handbook of Corpus Linguistics* (pp. 647–659). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_26

Grin, F., & Fürst, G. (2022). Measuring Linguistic Diversity: A Multi-level Metric. *Social Indicators Research*, *164*(2), 601–621. https://doi.org/10.1007/s11205-022-02934-5

Harris, Z. S. (1951). *Structural Linguistics*. The University of Chicago Press.

Harris, Z. S. (1957). Co-Occurrence and Transformation in Linguistic Structure. *Language*, *33*(3), 283. https://doi.org/10.2307/411155

Harris, Z. S. (1970). Linguistic Transformations for Information Retrieval. In *Papers in Structural and Transformational Linguistics* (pp. 458–471). Springer Netherlands. https://doi.org/10.1007/978-94-017-6059-1_24

Hatav, G. (1989). Aspects, *Aktionsarten,* and the time line. *Ling*, *27*(3), 487–516. https://doi.org/10.1515/ling.1989.27.3.487

Hellan, L. (2023). Unification and selection in Light Verb Constructions. A study of Norwegian. In *Light Verb Constructions as Complex Verbs* (pp. 45–70). De Gruyter. https://doi.org/10.1515/9783110747997-003

Hinrichs, E. W. (1985). *A Compositional Semantics for Aktionsarten And Np Reference In English (Aspect, Montague Grammar, Events, Mass Terms, Bare Plurals)*. The Ohio State University.

Hovav, M.R., & Levin, B. (2015). The Syntax-Semantics Interface. In *the Handbook of Contemporary Semantic Theory* (pp. 593–624). Wiley. https://doi.org/10.1002/9781118882139.ch19

Hopper, P. J., & Thompson, S. A. (1980). Transitivity in Grammar and Discourse. *Language, 56*(2), 251–299. https://doi.org/10.1353/lan.1980.0017

Hrenek, É. (2021). Synonymy in light verb constructions of the type feledésbe + verb | Szinonímia a feledésbe + ige típusú funkcióigés szerkezetek körében. *Magyar Nyelvor, 145*(3), 277–311. https://doi.org/10.38143/NYR.2021.3.277

Huang, C.-R., Lin, J., Jiang, M., & Xu, H. (2014). Corpus-based Study and Identification of Mandarin Chinese Light Verb Variations. *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects,* 1–10. https://doi.org/10.3115/v1/W14-5301

Hüning, M., & Schlücker, B. (2015). Multi-word expressions. *Word-Formation: An International Handbook of the Languages of Europe*, *1*, 450–467. https://doi.org/10.1515/9783110246254-026/HTML

Hwang, W., Lin, C., & Shen, T. (2015). Good–Turing frequency estimation in a finite population. *Biometrical Journal*, *57*(2), 321–339. https://doi.org/10.1002/bimj.201300168

Imseng, D., Bourlard, H., & Garner, P. N. (2012). Using KL-divergence and multilingual information to improve ASR for under-resourced languages. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4869–4872. https://doi.org/10.1109/ICASSP.2012.6289010

Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity. *Language Learning*, *63*(s1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*, *34*(4), 537–553. https://doi.org/10.1177/0265532217710632

Jarvis, S., & Hashimoto, B. J. (2021). How operationalizations of word types affect measures of lexical diversity. *International Journal of Learner Corpus Research*, *7*(1), 163–194. https://doi.org/10.1075/ijlcr.20004.jar

Jespersen, O. (1924). *The Philosophy of Grammar.* George Allen & Unwin LTD.

Jespersen, O. (1965). *A Modern English Grammar on Historical Principles, Part VI, Morphology*. George Allen and Unwin Ltd.

Jespersen, O. (2003). *Essentials of English Grammar*. Routledge. https://doi.org/10.4324/9780203425787

Jespersen, O. (2010). *Selected Writings of Otto Jespersen (Routledge Revivals).* Routledge. https://doi.org/10.4324/9780203857199

Jespersen, O. (2013a). *A Modern English Grammar on Historical Principles-Volume 4, Syntax*. Routledge. https://doi.org/10.4324/9780203715956

Jespersen, O. (2013b). *Novial Lexike.* Routledge. https://doi.org/10.4324/9780203716021

Jespersen, O. (2013c). *Progress in Language, with special reference to English*. Routledge. https://doi.org/10.4324/9780203716076

Jespersen, O. (2013d). *The Philosophy of Grammar.* Routledge. https://doi.org/10.4324/9780203716045

Jespersen, O. (2013e). *A Modern English Grammar on Historical Principles - Volume 2, Syntax* (Reprint Version). Routledge. https://doi.org/10.4324/9780203715932

Jespersen, O. (2015). *Linguistica.* Routledge. https://doi.org/10.4324/9781315694368

Ježek, E. (2023). Semantic Co-composition in Light Verb Constructions. In *Light Verb Constructions as Complex Verbs* (pp. 221–238). De Gruyter. https://doi.org/10.1515/9783110747997-008

Jurafsky, D., & Martin, J. H. (2009). *Naïve bayes classifier approach to Word sense disambiguation.* Computational Lexical Semantics.

Jost, L. (2006). Entropy and diversity. *Oikos*, *113*(2), 363–375. https://doi.org/10.1111/j.2006.0030-1299.14714.x

Karimi-Doostan, G. (2005). Light verbs and structural case. *Lingua*, *115*(12), 1737–1756. https://doi.org/10.1016/j.lingua.2004.08.002

Kettnerová, V. (2023). Valency structure of complex predicates with Light Verbs. In *Light Verb Constructions as Complex Verbs* (pp. 19–44). De Gruyter. https://doi.org/10.1515/9783110747997-002

Keylock, C. J. (2005). Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy. *Oikos*, *109*(1), 203–207. https://doi.org/10.1111/j.0030-1299.2005.13735.x

Kilgarriff, A. (1997). *Using Word Frequency Lists to Measure Corpus Homogeneity and Similarity between Corpora*.

Kilgarriff, A. (2001). *Comparing Corpora. International Journal of Corpus Linguistics*, 6(1), 97–133. https://doi.org/10.1075/ijcl.6.1.05kil

Kilgarriff, A., & Rose, T. (1998). Measures for corpus similarity and homogeneity. In *Proceedings of the Third Conference on Empirical Methods for Natural Language Processing,* 46–52. https://aclanthology.org/W98-1506.pdf

Kilgarriff, A., Baisa, V., Bušta, J. (2014). The Sketch Engine: ten years on. *Lexicography ASIALEX 1*, 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell. D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*: 105-11.

Kittilä, S. (2006). The anomaly of the verb "give" explained by its high (formal and semantic) transitivity. *Linguistics*, *44*(3), 569–612. https://doi.org/10.1515/LING.2006.019

Koerner, E. F. (2008). Hermann Paul and general linguistic theory. *Language Sciences*, *30*(1), 102-132.

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, *40*(4), 554–564. https://doi.org/10.1016/j.system.2012.10.012

Köhler, R., Altmann, G., & Piotrowski, R. (2005). *Quantitative Linguistik*. Walter de Gruyter. https://doi.org/10.1515/9783110155785

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the Validity of Lexical Diversity Indices Using Direct Judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and thought* (pp. 202–251). Cambridge University Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Langacker, R. W. (2014). Conceptualization, symbolization, and grammar. In *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure, Volume I Classic Edition*.

Langacker, R. W. (2020). Trees, assemblies, chains, and windows. *Constructions and Frames, 12*(1), 8–55. https://doi.org/10.1075/cf.00034.lan

Langacker, R. W. (2022). What could be more fundamental? In Karolina Krawczak, Barbara Lewandowska-Tomaszczyk, & Marcin Grygiel (Eds.), *Analogy and Contrast in Language: Perspectives from Cognitive Linguistics* (pp. 15–46). John Benjamins Publishing Company. https://doi.org/10.1075/hcp.73.01lan

Leipzig Corpora Collection: Indonesian mixed corpus based on material from 2013. *Leipzig Corpora Collection*. Dataset. https://corpora.uni-leipzig.de?corpusId=ind_mixed_2013.

Leski, J. M., & Kotas, M. P. (2018). Linguistically Defined Clustering of Data. International *Journal of Applied Mathematics and Computer Science, 28*(3), 545–557. https://doi.org/10.2478/amcs-2018-0042

Levin, B., & Hovav, M. R. (2017). Morphology and Lexical Semantics. In *the Handbook of Morphology* (pp. 248–271). Wiley. https://doi.org/10.1002/9781405166348.ch12

Lieber, R. (2004). *Morphology and Lexical Semantics*. Cambridge University Press.

Lieber, R. (2006). Syntax of Words. In *Encyclopedia of Language & Linguistics* (pp. 405–408). Elsevier. https://doi.org/10.1016/B0-08-044854-2/00144-9

Lieber, R. (2007). The category of roots and the roots of categories: what we learn from selection in derivation. *Morphology*, *16*(2), 247–272. https://doi.org/10.1007/s11525-006-9106-2

Lieber, R. (2010). *On the lexical semantics of compounds* (pp. 127–144). https://doi.org/10.1075/cilt.311.11lie

Lieber, R. (2011). *A Lexical Semantic Approach to Compounding*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199695720.013.0005

Lieber, R., & Štekauer, P. (2011). *Introduction: Status and Definition of Compounding*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199695720.013.0001

Lin, J., Xu, H., Jiang, M., & Huang, C. R. (2014). Annotation and classification of light verbs and light verb variations in Mandarin Chinese. In *Proceedings of workshop on lexical and grammatical resources for language processing* (pp. 75-82).

Liu, H., & Li, W. (2010). Language clusters based on linguistic complex networks. *Chinese Science Bulletin, 55*(30), 3458–3465. https://doi.org/10.1007/s11434-010-4114-3

Magueresse, A., Carles, V., and Heetderks, E. (2020). Low-resource Languages: A Review of Past Work and Future Challenges. *Computation and Language*. https://doi.org/10.48550/arXiv.2006.07264

Manin, D. Y. (2009). Mandelbrot's Model for Zipf's Law: Can Mandelbrot's Model Explain Zipf's Law for Language? *Journal of Quantitative Linguistics, 16*(3), 274–285. https://doi.org/10.1080/09296170902850358

Marantz, A. (2013). Verbal argument structure: Events and participants. *Lingua*, *130*, 152–168. https://doi.org/10.1016/J.LINGUA.2012.10.012

Mahowald, K., Jurafsky, D., & Norris, M. (2021). Concord begets concord: A Bayesian model of nominal concord typology. *Proceedings of the Linguistic Society of America*, *6*(1), 541. https://doi.org/10.3765/plsa.v6i1.4988

Malchukov, A. L. (2006). Transitivity parameters and transitivity alternations. In *Case, Valency and Transitivity* (pp. 329–357). https://doi.org/10.1075/slcs.77.21mal

Mastrofini, R. (2023). When lightness meets lexical aspect. A corpus-based account of English Light Verb Extensions. In *Light Verb Constructions as Complex Verbs* (pp. 201–218). De Gruyter. https://doi.org/10.1515/9783110747997-007

Mattissen, J. (2023). Light Verbs and 'light nouns' in polysynthetic languages. In *Light Verb Constructions as Complex Verbs* (pp. 275–304). De Gruyter. https://doi.org/10.1515/9783110747997-011

Maxwelll-Smith, Z., Kohler, M., & Suominen, H. (2022). Scoping natural language processing in Indonesian and Malay for education applications. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 171–228. https://doi.org/10.18653/v1/2022.acl-srw.15

McCarthy, D., Apidianaki, M., & Erk, K. (2016). Word Sense Clustering and Clusterability. *Computational Linguistics, 42*(2), 245–275. https://doi.org/10.1162/COLI_a_00247

McEnery, T., & Brezina, V. (2022). *Fundamental Principles of Corpus Linguistics*. Cambridge University Press. https://doi.org/10.1017/9781107110625

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

Mehl, S. (2019). Light verb semantics in the international corpus of English: Onomasiological variation, identity evidence and degrees of lightness. *English Language and Linguistics, 23*(1), 55–80. https://doi.org/10.1017/S1360674317000302

Mel'čuk, I. (2006). *Aspects of the Theory of Morphology*. Mouton de Gruyter.

Mel'čuk, I. (2022). Support (=Light) Verbs. *Neophilologica*, *34*, 1–30. https://doi.org/10.31261/NEO.2022.34.03

Miyamoto, T. (2000). *The Light Verb Construction in Japanese (Vol. 29)*. John Benjamins Publishing Company. https://doi.org/10.1075/la.29

Miyamoto, T., & Kishimoto, H. (2016). Light verb constructions with verbal nouns. In *Handbook of Japanese Lexicon and Word Formation* (pp. 425–458). De Gruyter. https://doi.org/10.1515/9781614512097-016

Mlac, E., & Tournadre, N. (2021). The semantics of the verb give in Tibetan: The development of the transfer construction and the honorific domain. In *Constructional Approaches to Language* (Vol. 29). https://doi.org/10.1075/cal.29.07eri

Moisl, H. (2010). Variable scaling in cluster analysis of linguistic data. *Corpus Linguistics and Linguistic Theory, 6*(1). https://doi.org/10.1515/cllt.2010.004

Moisl, H. (2015). *Cluster Analysis for Corpus Linguistics.* De Gruyter. https://doi.org/10.1515/9783110363814

Moisl, H. (2021). *Topological mapping for visualisation of high-dimensional historical linguistic data* (pp. 209–224). https://doi.org/10.1075/cilt.356.14moi

Mudraya, O., Piao, S. S. L., Rayson, P., Sharoff, S., Babych, B., & Löfberg, L. (2008). Automatic extraction of translation equivalents of phrasal and light verbs in English and Russian. In *Phraseology: An Interdisciplinary Perspective*. https://doi.org/10.1075/z.139.25mud

Nagy T., István, Rácz, A., & Vincze, V. (2020). Detecting light verb constructions across languages. *Natural Language Engineering*, *26*(3), 319–348. https://doi.org/10.1017/S1351324919000330

Naeem, S., & Wumaier, A. (2018). Study and Implementing K-mean Clustering Algorithm on English Text and Techniques to Find the Optimal Value of K. *International Journal of Computer Applications, 182*(31), 7–14. https://doi.org/10.5120/ijca2018918234

Nugraha, D. S. (2022). Identifying Light Verb Constructions in Indonesian: A Direct Translation Approach. *International Journal of Language and Literary Studies*, *4*(3), 298–311. https://doi.org/10.36892/ijlls.v4i3.1042

Nugraha, D. S. (2023a). Morphosemantic Features of Membuat 'Make' in The Light Verb Constructions of Indonesian. *LiNGUA: Jurnal Ilmu Bahasa Dan Sastra*, *17*(2), 131–142. https://doi.org/10.18860/ling.v17i2.17757

Nugraha, D. S. (2023b). Morphosemantic Features of Mengambil 'Take' in the Light Verb Constructions of Indonesian. *International Journal of Linguistics and Translation Studies*, *4*(3), 120–138. https://doi.org/10.36892/ijlts.v4i3.327

Nugraha, D. S. (2023c). Morphosyntactic Features of Membuat 'Make' in the Light Verb Constructions of Indonesian. *European Journal of Language and Culture Studies*, *2*(2), 33–43. https://doi.org/10.24018/ejlang.2023.2.2.80

Nugraha, D. S. (2024a). A Corpus-Based Study of Memberi 'Give' Light Verb Constructions. *International Journal of Society, Culture and Language,* *12*(2), 104–120. https://doi.org/10.22034/ijscl.2024.2023112.3389

Nugraha, D. S. (2024b). Morphosemantic Features of Memenuhi 'Meet' in The Light Verb Constructions of Indonesian. *Linguistik Indonesia,* *42*(2), 461–480. https://doi.org/10.26499/li.v42i2.638

Nugraha, D. S. (2024c). Analyzing Prefix /me(N)-/ in the Indonesian Affixation: A Corpus-Based Morphology. *Theory and Practice in Language Studies,* *14*(6), 1697–1711. https://doi.org/10.17507/tpls.1406.10

Nugraha, D. S. (2024d). A Morphological Analysis of the Indonesian Suffixation: A Look at the Different Types of Affixes and Their Semantic Changes. *GEMA: Journal of Language Studies,* *24*(4), 1–24. http://doi.org/10.17576/gema-2024-2404-

Nugraha, D. S., & Vincze, V. (2024). Towards an Empirical Understanding of membawa 'bring': Corpus Insights into Indonesian Light Verb Constructions. *Jurnal Arbitrer*, 11(3), 278–296. https://doi.org/10.25077/ar.11.3.278-296.2024

Oakes, M. (2019). *Statistics for Corpus Linguistics*. Edinburgh University Press. https://doi.org/10.1515/9781474471381

Ohnheiser, I. (2015). Compounds and multi-word expressions in slavic. *Word-Formation: An International Handbook of the Languages of Europe*, *1*, 757–779. https://doi.org/10.1515/9783110246254-045/HTML

Ong, C. S. B., & Rahim, H. A. (2021). Analysis of light verb construction use in L1 and L2: Insights from british and Malaysian student writing. *TESL-EJ*, *25*(2).

Özge, D., Ünal, G., & Bayırlı, İ. K. (2022). Assigning meaning to light verbs in Turkish | Türkçede katkısız eylemlere anlam atanması. *Dilbilim Arastirmalari Dergisi*, *33*(1), 1–27. https://doi.org/10.18492/dad.917075

Perdue, C. (2006). "Creating language anew": some remarks on an idea of Bernard Comrie's. *Linguistics*, *44*(4). https://doi.org/10.1515/LING.2006.027

Piątkowski, Ł. (2019). On corpus-based teaching of light verb constructions in German during foreign language lessons. Syntagmatic patterns as "instructions for use" for learners of German | Zur korpusbasierten Vermittlung der Funktionsverbgefüge im DaF-Unterricht. Syntagmatisc. *Glottodidactica*, *46*(1), 127–144. https://doi.org/10.14746/gl.2019.46.1.08

Pilgrim, C., Guo, W., & Hills, T. T. (2024). The rising entropy of English in the attention economy. *Communications Psychology*, *2*(1), 70. https://doi.org/10.1038/s44271-024-00117-1

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review, 21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

Pompei, A. (2023). How light is "give" as a Light Verb? A case study on the actionality of Latin Light Verb Constructions (with some references to Romance languages). In *Light Verb Constructions as Complex Verbs: Features, Typology and Function*. https://doi.org/10.1515/9783110747997-006

Pompei, A., & Piunno, V. (2023). Light Verb Constructions in Romance languages: An attempt to explain systematic irregularity. In *Light Verb Constructions as Complex Verbs: Features, Typology and Function*. https://doi.org/10.1515/9783110747997-005

Popescu, I.-I., Altmann, G., & Köhler, R. (2010). Zipf's law—another view. *Quality & Quantity, 44*(4), 713–731. https://doi.org/10.1007/s11135-009-9234-y

Prokhorov, V., Shareghi, E., Li, Y., Pilehvar, M. T., & Collier, N. (2019). On the Importance of the Kullback-Leibler Divergence Term in Variational Autoencoders for Text Generation. *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 118–127. https://doi.org/10.18653/v1/D19-5612

Raschka, S. (2014). *Naive Bayes and Text Classification I - Introduction and Theory*. https://doi.org/https://doi.org/10.48550/arXiv.1410.5329

Reis, M. (1978). Hermann Paul. *Beiträge Zur Geschichte Der Deutschen Sprache Und Literatur (PBB)*, *1978*(100). https://doi.org/10.1515/bgsl.1978.1978.100.159

Ricca, D. (2015). Verb-noun compounds in romance. *Word-Formation: An International Handbook of the Languages of Europe*, *1*, 688–707. https://doi.org/10.1515/9783110246254-041/HTML

Rinaldo, A., & Wasserman, L. (2010). Generalized density clustering. *The Annals of Statistics, 38*(5). https://doi.org/10.1214/10-AOS797

Ronan, P. (2014). Light verb constructions in the history of English. In *Studies in Corpus Linguistics* (Vol. 63). https://doi.org/10.1075/scl.63.05ron

Ronan, P. (2019). Simple versus Light Verb Constructions in Late Modern Irish English Correspondence: A Qualitative and Quantitative Analysis. *Studia Neophilologica*, *91*(1), 31–48. https://doi.org/10.1080/00393274.2019.1578182

Ronan, P., & Schneider, G. (2015). Determining light verb constructions in contemporary British and Irish English. *International Journal of Corpus Linguistics, 20*(3), 326–354. https://doi.org/10.1075/ijcl.20.3.03ron

Ross, A. S. C., & Herdan, G. (1960). Type-Token Mathematics: A Textbook of Mathematical Linguistics. *Journal of the Royal Statistical Society. Series A (General)*, *123*(3), 341. https://doi.org/10.2307/2342480

Rosell, M. (2009). Text Clustering Exploration: Swedish Text Representation and Clustering Results Unraveled [PhD Dissertation. KTH School of Computer Science and Communication]. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A209282&dswid=1882

Roßdeutscher, A. (2014). When roots license and when they respect semantico-syntactic structure in verbs. In *the Syntax of Roots and the Roots of Syntax* (pp. 282–309). Oxford University Press Oxford. https://doi.org/10.1093/acprof:oso/9780199665266.003.0013

Rouhi, A., & Heidari, A. (2015). Light verb Constructions in Azeri-Turkish/Persian Code-switching based on the Matrix Language Frame Model. *Language Related Research*, *6*(1), 111–129. https://lrr.modares.ac.ir/article-14-12233-en.html

Salido, M. G., & Garcia, M. (2023). On the unpredictability of Support Verbs: A distributional study of Spanish tomar. In *Light Verb Constructions as Complex Verbs: Features, Typology and Function* (pp. 239–255). De Gruyter. https://doi.org/10.1515/9783110747997-009

Sebastian, D., Purnomo, H. D., & Sembiring, I. (2022). BERT for Natural Language Processing in Bahasa Indonesia. *2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, 204–209. https://doi.org/10.1109/ICICyTA57421.2022.10038230

Serra-Peralta, M., Serrà, J., & Corral, Á. (2021). Heaps' law and vocabulary richness in the history of classical music harmony. *EPJ Data Science*, *10*(1), 40. https://doi.org/10.1140/epjds/s13688-021-00293-8

Sheather, S. J. (2004). Density Estimation. *Statistical Science, 19*(4). https://doi.org/10.1214/088342304000000297

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shannon, C. E. (1951). Prediction and Entropy of Printed English. *Bell System Technical Journal*, *30*(1), 50–64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x

Sheng, D. (2023). Statistics in corpus linguistics: a practical guide. *Social Semiotics, 33*(4), 909–912. https://doi.org/10.1080/10350330.2021.1969215

Si, F. (2021). Towards a cartography of light verbs. In *Linguistik Aktuell* (Vol. 267). https://doi.org/10.1075/la.267.10si

Silverman, B. W. (2018). *Density Estimation for Statistics and Data Analysis*. Routledge. https://doi.org/10.1201/9781315140919

Simpson, E. H. (1949). Measurement of Diversity. *Nature*, *163*(4148), 688–688. https://doi.org/10.1038/163688a0

Singleton, D. (2016). *Language and the Lexicon.* Routledge. https://doi.org/10.4324/9781315824796

Singh, A.K. (2008). Natural language processing for less privileged languages: Where do we come from? where are we going? In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Snider, T. (2021). Light Verbs in Biblical Hebrew. In *Studies in Semitic Languages and Linguistics* (Vol. 102). https://doi.org/10.1163/9789004448858_010

Strohbach, M. (1984). *Johann Christoph Adelung*. De Gruyter. https://doi.org/10.1515/9783110852493

Suzgun, M., Melas-Kyriazi, L., & Jurafsky, D. (2022). *Follow the Wisdom of the Crowd: Effective Text Generation via Minimum Bayes Risk Decoding*.

Tadmor, U. (2018). Malay-Indonesian. In B. Comrie (Ed.), *The World's Major Languages 3rd Edition*. Routledge.

Taylor, J. R. (2015). Prototype Theory in Linguistics. In *International Encyclopedia of the Social &amp; Behavioral Sciences: Second Edition.* https://doi.org/10.1016/B978-0-08-097086-8.57034-2

Thinsungnoen, T., Kaoungku, N., Durongdumronchai, P., Kerdprasop, K., & Kerdprasop, N. (2015). The Clustering Validity with Silhouette and Sum of Squared Errors. *The Proceedings of the 2nd International Conference on Industrial Application Engineering 2015*, 44–51. https://doi.org/10.12792/iciae2015.012

Toluspayeva, D., Shubina, O., Apshe, K., Shukan, A., Nurguzhina, G., & Mirambaevna, A. B. (2024). Affixation in Morphological Word Formation and Construction of Lexemes in the English and Kazakh Languages. *International Journal of Society, Culture & Language, 12*(1), 326–336. https://doi.org/10.22034/ijscl.2024.2019058.3318

Tsou, B. K., & Yip, K.-F. (2020). A corpus-based comparative study of light verbs in three Chinese speech communities. In M. Le Nguyen, M. C. Luong, & S. Song (Eds.), *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation* (pp. 302–311). Association for Computational Linguistics. https://aclanthology.org/2020.paclic-1.35/

Tsvetkov, Y. (2017). *Opportunities and challenges in working with low-resource languages*. Carnegie Mellon University.

Vincze, V. (2011). *Semi-Compositional Noun + Verb Constructions: Theoretical Questions and Computational Linguistic Analyses*. University of Szeged.

Vincze, V. (2014). Valency Frames in a Hungarian Corpus. *Journal of Quantitative Linguistics, 21*(2), 153–176. https://doi.org/10.1080/09296174.2014.882188

Vincze, V., István Nagy, T., & Zsibrita, J. (2013). Learning to detect English and Hungarian light verb constructions. *ACM Transactions on Speech and Language Processing, 10*(2). https://doi.org/10.1145/2483691.2483695

Wallis, S. (2021). *Statistics in Corpus Linguistics Research*. Routledge.

Wang, C.-A. A., & Wu, H.-H. I. (2020). Light verbs in verbal reduplication. *Studia Linguistica*, *74*(2), 337–359. https://doi.org/10.1111/stul.12128

Wang, D., Jiang, G., & Zheng, Y. (2023). Walking out of the light verb jungle: Exploring the translation strategies of light verb constructions in Chinese–English consecutive interpreting. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1113973

Wasserman, L. (2004). *All of Statistics.* Springer New York. https://doi.org/10.1007/978-0-387-21736-9

Wasserman, L. (2006). *All of Nonparametric Statistics.* Springer New York. https://doi.org/10.1007/0-387-30623-4

Węglarczyk, S. (2018). Kernel density estimation and its application. *ITM Web of Conferences, 23, 00037*. https://doi.org/10.1051/itmconf/20182300037

Wittenberg, E., & Piñango, M. M. (2011). Processing light verb constructions. *Mental Lexicon*, *6*(3), 393–413. https://doi.org/10.1075/ml.6.3.03wit

Wu, T., Tao, C., Wang, J., Yang, R., Zhao, Z., & Wong, N. (2025). Rethinking Kullback-Leibler Divergence in Knowledge Distillation for Large Language Models. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. Di Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 5737–5755). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.383/

Xu, H., Jiang, M., Lin, J., & Huang, C.-R. (2022). Light verb variations and varieties of Mandarin Chinese: Comparable corpus driven approaches to grammatical variations. *Corpus Linguistics and Linguistic Theory, 18*(1), 145–173. https://doi.org/10.1515/cllt-2019-0049

Yamano, T. (2001). A Possible Extension of Shannon’s Information Theory. *Entropy*, *3*(4), 280–292. https://doi.org/10.3390/e3040280

Yim, C. (2020). Light verb ellipsis constructions in Korean. *Language and Linguistics*, *21*(2), 351–374. https://doi.org/10.1075/lali.00064.yim

Zhang, Y., Shan, S., Yan, C., Qiu, J., & Xiong, L. (2016). Herdan-Heaps Law Corresponds to Lotka’s Law: A Dynamic Perspective from Simon’s Model. *Journal of Quantitative Linguistics*, *23*(2), 211–227. https://doi.org/10.1080/09296174.2016.1142329

Zipf, G. K. (2013). *The Psycho-Biology of Language.* Routledge. https://doi.org/10.4324/9781315009421

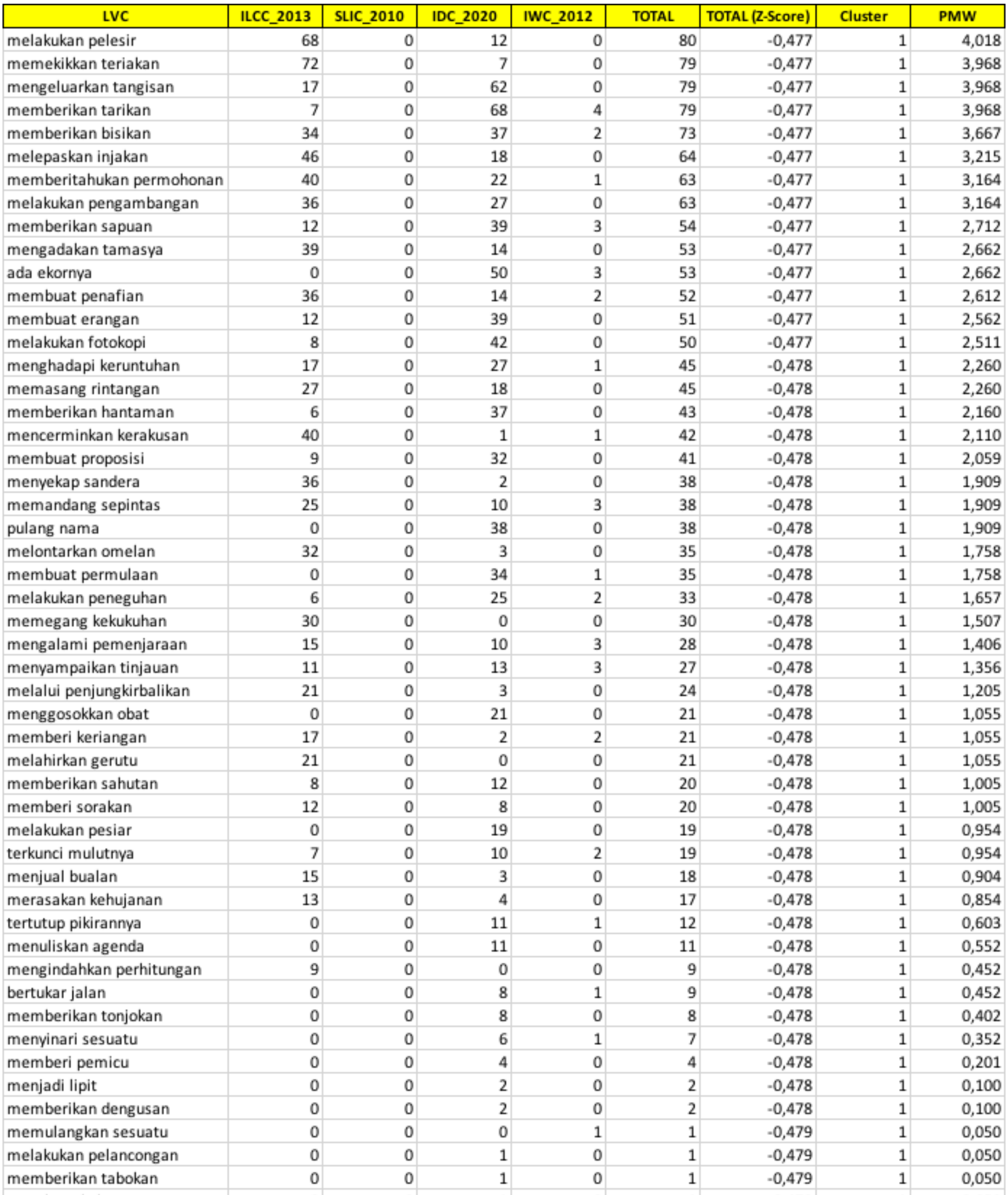
Appendices

Appendix A

Sample list of the Indonesian LVCs

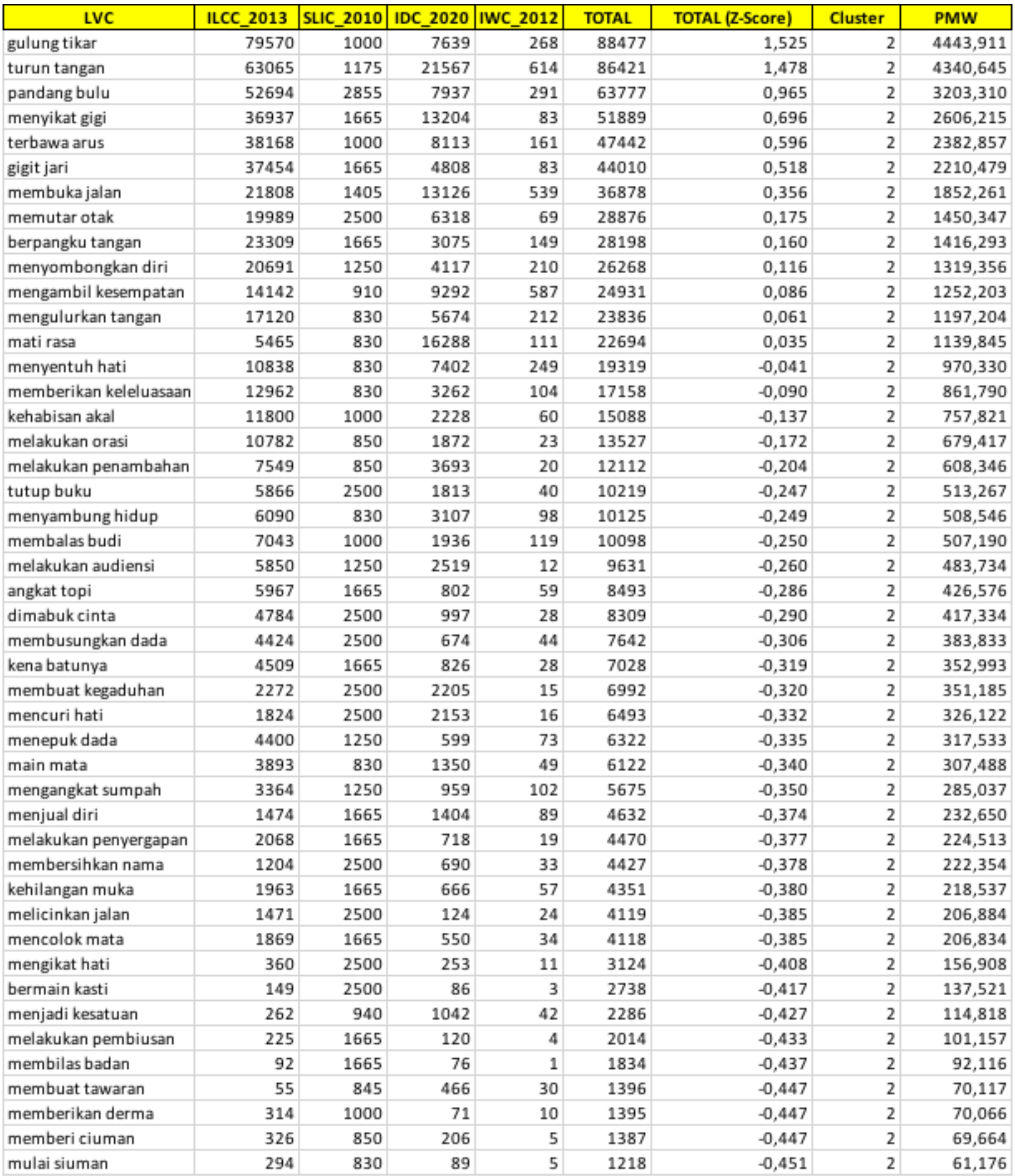
The complete list of Indonesian LVCs are provided in the external repository. Please pay a visit to: https://github.com/dsnugrahaCL/LVCs_Indonesian.

A.1 Sample list of the low-frequency LVCs



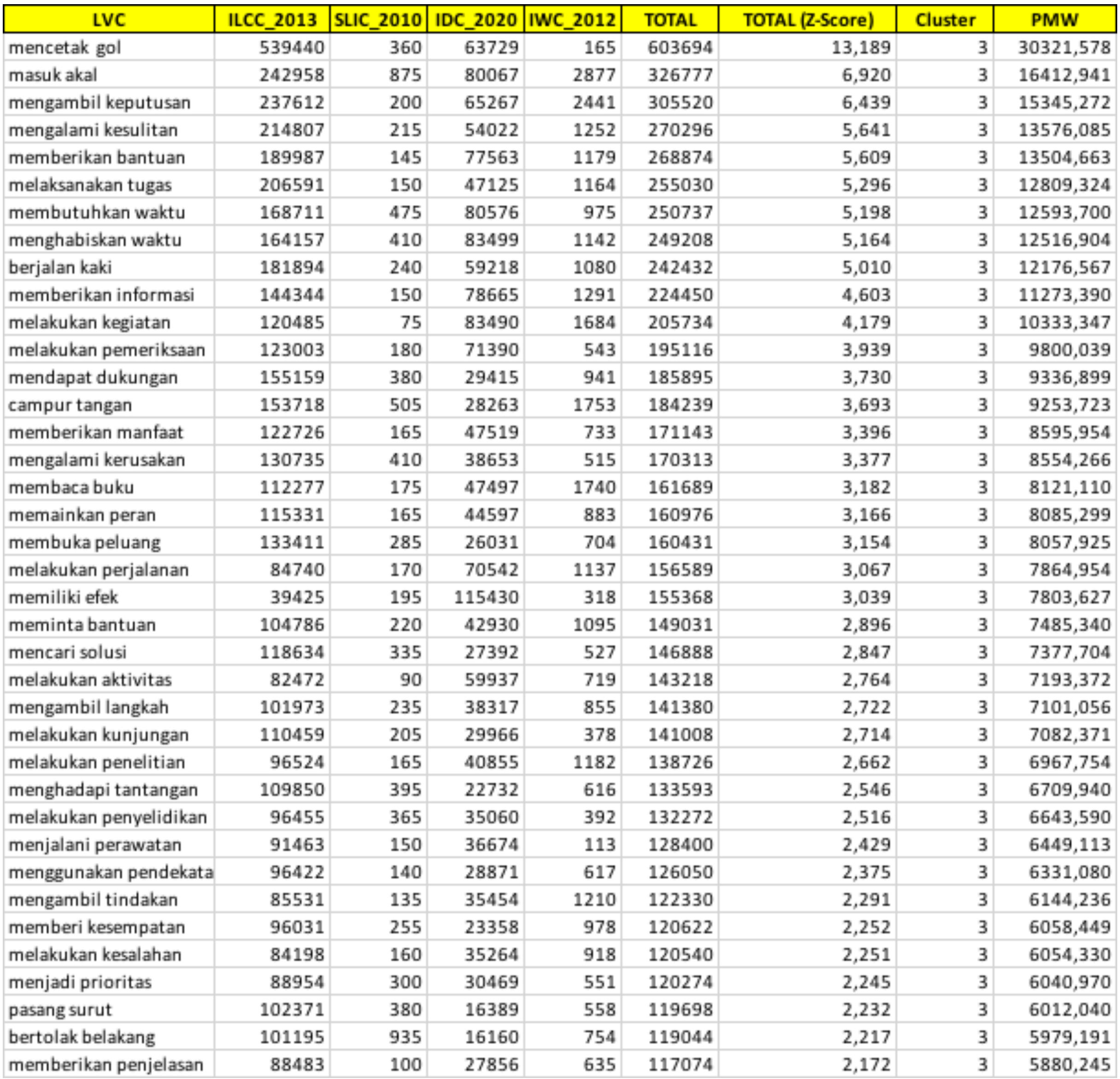
LVC	ILCC_2013	SLIC_2010	IDC_2020	IWC_2012	TOTAL	TOTAL (Z-Score)	Cluster	PMW
melakukan pelesir	68	0	12	0	80	-0,477	1	4,018
memekikkan teriakan	72	0	7	0	79	-0,477	1	3,968
mengeluarkan tangisan	17	0	62	0	79	-0,477	1	3,968
memberikan tarikan	7	0	68	4	79	-0,477	1	3,968
memberikan bisikan	34	0	37	2	73	-0,477	1	3,667
melepaskan injakan	46	0	18	0	64	-0,477	1	3,215
memberitahukan permohonan	40	0	22	1	63	-0,477	1	3,164
melakukan pengambangan	36	0	27	0	63	-0,477	1	3,164
memberikan sapuan	12	0	39	3	54	-0,477	1	2,712
mengadakan tamasya	39	0	14	0	53	-0,477	1	2,662
ada ekornya	0	0	50	3	53	-0,477	1	2,662
membuat penafian	36	0	14	2	52	-0,477	1	2,612
membuat erangan	12	0	39	0	51	-0,477	1	2,562
melakukan fotokopi	8	0	42	0	50	-0,477	1	2,511
menghadapi keruntuhan	17	0	27	1	45	-0,478	1	2,260
memasang rintangan	27	0	18	0	45	-0,478	1	2,260
memberikan hantaman	6	0	37	0	43	-0,478	1	2,160
mencerminkan kerakusan	40	0	1	1	42	-0,478	1	2,110
membuat proposisi	9	0	32	0	41	-0,478	1	2,059
menyekap sandera	36	0	2	0	38	-0,478	1	1,909
memandang sepintas	25	0	10	3	38	-0,478	1	1,909
pulang nama	0	0	38	0	38	-0,478	1	1,909
melontarkan omelan	32	0	3	0	35	-0,478	1	1,758
membuat permulaan	0	0	34	1	35	-0,478	1	1,758
melakukan peneguhan	6	0	25	2	33	-0,478	1	1,657
memegang kekukuhan	30	0	0	0	30	-0,478	1	1,507
mengalami pemenjaraan	15	0	10	3	28	-0,478	1	1,406
menyampaikan tinjauan	11	0	13	3	27	-0,478	1	1,356
melalui penjungkirbalikan	21	0	3	0	24	-0,478	1	1,205
menggosokkan obat	0	0	21	0	21	-0,478	1	1,055
memberi keriangan	17	0	2	2	21	-0,478	1	1,055
melahirkan gerutu	21	0	0	0	21	-0,478	1	1,055
memberikan sahutan	8	0	12	0	20	-0,478	1	1,005
memberi sorakan	12	0	8	0	20	-0,478	1	1,005
melakukan pesiar	0	0	19	0	19	-0,478	1	0,954
terkunci mulutnya	7	0	10	2	19	-0,478	1	0,954
menjual bualan	15	0	3	0	18	-0,478	1	0,904
merasakan kehujanan	13	0	4	0	17	-0,478	1	0,854
tertutup pikirannya	0	0	11	1	12	-0,478	1	0,603
menuliskan agenda	0	0	11	0	11	-0,478	1	0,552
mengindahkan perhitungan	9	0	0	0	9	-0,478	1	0,452
bertukar jalan	0	0	8	1	9	-0,478	1	0,452
memberikan tonjokan	0	0	8	0	8	-0,478	1	0,402
menyinari sesuatu	0	0	6	1	7	-0,478	1	0,352
memberi pemicu	0	0	4	0	4	-0,478	1	0,201
menjadi lipit	0	0	2	0	2	-0,478	1	0,100
memberikan dengusan	0	0	2	0	2	-0,478	1	0,100
memulangkan sesuatu	0	0	0	1	1	-0,479	1	0,050
melakukan pelancongan	0	0	1	0	1	-0,479	1	0,050
memberikan tabokan	0	0	1	0	1	-0,479	1	0,050

A.2 Sample list of the medium-frequency LVCs



LVC	ILCC_2013	SLIC_2010	IDC_2020	IWC_2012	TOTAL	TOTAL (Z-Score)	Cluster	PMW
gulung tikar	79570	1000	7639	268	88477	1,525	2	4443,911
turun tangan	63065	1175	21567	614	86421	1,478	2	4340,645
pandang bulu	52694	2855	7937	291	63777	0,965	2	3203,310
menyikat gigi	36937	1665	13204	83	51889	0,696	2	2606,215
terbawa arus	38168	1000	8113	161	47442	0,596	2	2382,857
gigit jari	37454	1665	4808	83	44010	0,518	2	2210,479
membuka jalan	21808	1405	13126	539	36878	0,356	2	1852,261
memutar otak	19989	2500	6318	69	28876	0,175	2	1450,347
berpangku tangan	23309	1665	3075	149	28198	0,160	2	1416,293
menyombongkan diri	20691	1250	4117	210	26268	0,116	2	1319,356
mengambil kesempatan	14142	910	9292	587	24931	0,086	2	1252,203
mengulurkan tangan	17120	830	5674	212	23836	0,061	2	1197,204
mati rasa	5465	830	16288	111	22694	0,035	2	1139,845
menyentuh hati	10838	830	7402	249	19319	-0,041	2	970,330
memberikan keleluasaan	12962	830	3262	104	17158	-0,090	2	861,790
kehabisan akal	11800	1000	2228	60	15088	-0,137	2	757,821
melakukan orasi	10782	850	1872	23	13527	-0,172	2	679,417
melakukan penambahan	7549	850	3693	20	12112	-0,204	2	608,346
tutup buku	5866	2500	1813	40	10219	-0,247	2	513,267
menyambung hidup	6090	830	3107	98	10125	-0,249	2	508,546
membalas budi	7043	1000	1936	119	10098	-0,250	2	507,190
melakukan audiensi	5850	1250	2519	12	9631	-0,260	2	483,734
angkat topi	5967	1665	802	59	8493	-0,286	2	426,576
dimabuk cinta	4784	2500	997	28	8309	-0,290	2	417,334
membusungkan dada	4424	2500	674	44	7642	-0,306	2	383,833
kena batunya	4509	1665	826	28	7028	-0,319	2	352,993
membuat kegaduhan	2272	2500	2205	15	6992	-0,320	2	351,185
mencuri hati	1824	2500	2153	16	6493	-0,332	2	326,122
menepuk dada	4400	1250	599	73	6322	-0,335	2	317,533
main mata	3893	830	1350	49	6122	-0,340	2	307,488
mengangkat sumpah	3364	1250	959	102	5675	-0,350	2	285,037
menjual diri	1474	1665	1404	89	4632	-0,374	2	232,650
melakukan penyergapan	2068	1665	718	19	4470	-0,377	2	224,513
membersihkan nama	1204	2500	690	33	4427	-0,378	2	222,354
kehilangan muka	1963	1665	666	57	4351	-0,380	2	218,537
melicinkan jalan	1471	2500	124	24	4119	-0,385	2	206,884
mencolok mata	1869	1665	550	34	4118	-0,385	2	206,834
mengikat hati	360	2500	253	11	3124	-0,408	2	156,908
bermain kasti	149	2500	86	3	2738	-0,417	2	137,521
menjadi kesatuan	262	940	1042	42	2286	-0,427	2	114,818
melakukan pembiusan	225	1665	120	4	2014	-0,433	2	101,157
membilas badan	92	1665	76	1	1834	-0,437	2	92,116
membuat tawaran	55	845	466	30	1396	-0,447	2	70,117
memberikan derma	314	1000	71	10	1395	-0,447	2	70,066
memberi ciuman	326	850	206	5	1387	-0,447	2	69,664
mulai siuman	294	830	89	5	1218	-0,451	2	61,176

A.3 Sample list of the high-frequency LVCs



LVC	ILCC_2013	SLIC_2010	IDC_2020	IWC_2012	TOTAL	TOTAL (Z-Score)	Cluster	PMW
mencetak gol	539440	360	63729	165	603694	13,189	3	30321,578
masuk akal	242958	875	80067	2877	326777	6,920	3	16412,941
mengambil keputusan	237612	200	65267	2441	305520	6,439	3	15345,272
mengalami kesulitan	214807	215	54022	1252	270296	5,641	3	13576,085
memberikan bantuan	189987	145	77563	1179	268874	5,609	3	13504,663
melaksanakan tugas	206591	150	47125	1164	255030	5,296	3	12809,324
membutuhkan waktu	168711	475	80576	975	250737	5,198	3	12593,700
menghabiskan waktu	164157	410	83499	1142	249208	5,164	3	12516,904
berjalan kaki	181894	240	59218	1080	242432	5,010	3	12176,567
memberikan informasi	144344	150	78665	1291	224450	4,603	3	11273,390
melakukan kegiatan	120485	75	83490	1684	205734	4,179	3	10333,347
melakukan pemeriksaan	123003	180	71390	543	195116	3,939	3	9800,039
mendapat dukungan	155159	380	29415	941	185895	3,730	3	9336,899
campur tangan	153718	505	28263	1753	184239	3,693	3	9253,723
memberikan manfaat	122726	165	47519	733	171143	3,396	3	8595,954
mengalami kerusakan	130735	410	38653	515	170313	3,377	3	8554,266
membaca buku	112277	175	47497	1740	161689	3,182	3	8121,110
memainkan peran	115331	165	44597	883	160976	3,166	3	8085,299
membuka peluang	133411	285	26031	704	160431	3,154	3	8057,925
melakukan perjalanan	84740	170	70542	1137	156589	3,067	3	7864,954
memiliki efek	39425	195	115430	318	155368	3,039	3	7803,627
meminta bantuan	104786	220	42930	1095	149031	2,896	3	7485,340
mencari solusi	118634	335	27392	527	146888	2,847	3	7377,704
melakukan aktivitas	82472	90	59937	719	143218	2,764	3	7193,372
mengambil langkah	101973	235	38317	855	141380	2,722	3	7101,056
melakukan kunjungan	110459	205	29966	378	141008	2,714	3	7082,371
melakukan penelitian	96524	165	40855	1182	138726	2,662	3	6967,754
menghadapi tantangan	109850	395	22732	616	133593	2,546	3	6709,940
melakukan penyelidikan	96455	365	35060	392	132272	2,516	3	6643,590
menjalani perawatan	91463	150	36674	113	128400	2,429	3	6449,113
menggunakan pendekata	96422	140	28871	617	126050	2,375	3	6331,080
mengambil tindakan	85531	135	35454	1210	122330	2,291	3	6144,236
memberi kesempatan	96031	255	23358	978	120622	2,252	3	6058,449
melakukan kesalahan	84198	160	35264	918	120540	2,251	3	6054,330
menjadi prioritas	88954	300	30469	551	120274	2,245	3	6040,970
pasang surut	102371	380	16389	558	119698	2,232	3	6012,040
bertolak belakang	101195	935	16160	754	119044	2,217	3	5979,191
memberikan penjelasan	88483	100	27856	635	117074	2,172	3	5880,245

Appendix B

List of the verb elements and their optimal pairing-rate within Indonesian LVCs

This list presents all verb elements identified in Indonesian LVCs datasets. The number following each verb indicates its optimal possibility of the distinctive combination with noun elements.

melakukan *do / carry out* 141
memberikan *give / provide* 120
membuat *make / create* 53
mengalami *experience* 34
memberi *give* 31
mengambil *take* 21
menjadi *become* 18
memiliki *have / possess* 16
menyampaikan *deliver / convey* 16
mengadakan *hold / organize* 12
menimbulkan *cause / generate* 12
membawa *bring / carry* 11
merasa *feel* 10
melontarkan *launch / hurl* 9
memasang *install / attach* 8
menderita *suffer* 8
melaksanakan *implement* 7
mendapatkan *obtain / get* 7
menghadapi *face / confront* 7
melayangkan *send / file* 6
membuka *open* 6
mendapat *get* 6
menjalani *undergo* 6
merasakan *feel / sense* 6
membangun *build / construct* 5
mencapai *achieve / reach* 5
mengajukan *submit / propose* 5
mengeluarkan *issue / release* 5
melepaskan *release / let go* 4
membutuhkan *need / require* 4
memicu *trigger / spark* 4
memulai *start / begin* 4
menaruh *place / put* 4
mencari *search / look for* 4
menjalankan *carry out / run* 4
menjelang *approaching* 4
menyebabkan *cause* 4
menyediakan *provide* 4
terjadi *happen / occur* 4
berjalan *walk / proceed* 3
bermain *play* 3
membersihkan *clean* 3
memegang *hold* 3
memperoleh *gain / acquire* 3
memukul *hit / strike* 3
menggelar *hold / stage* 3
menggunakan *use* 3
mengundang *invite* 3
meninggalkan *leave* 3
menjaga *guard / maintain* 3
menjatuhkan *drop / overthrow* 3
menunjukkan *show / indicate* 3
meraih *reach / achieve* 3
kehilangan *lose* 2
masuk *enter* 2
melahirkan *give birth* 2
melalui *go through / via* 2
melemparkan *throw* 2
melihat *see / look* 2
memainkan *play (an instrument)* 2
membalas *reply / retaliate* 2
membasuh *wash* 2
membawakan *bring (for)* 2
membentangkan *spread / unfold* 2
membentuk *form / shape* 2
membubuhkan *affix / add* 2
memunculkan *bring up / raise* 2
menarik *attract / pull* 2
menawarkan *offer* 2
menciptakan *create* 2
mencium *kiss / smell* 2
mencuri *steal* 2
menemukan *find / discover* 2
mengangkat *lift / appoint* 2
mengenai *about / regarding* 2
mengikat *bind / tie* 2
mengikuti *follow / attend* 2
menginjak *step on* 2
mengulurkan *extend / stretch* 2
menjual *sell* 2
menyalurkan *channel / distribute* 2
menyatakan *state / declare* 2
pergi *go* 2
ada *exist / be present* 1
adu *compete / argue* 1
angkat *lift / raise* 1

berbalik *turn around* 1
bergoyang *sway / shake* 1
berhutang *owe* 1
beristirahat *rest* 1
berlatih *practice* 1
berpangku *sit on lap* 1
bertemu *meet* 1
bertolak *depart* 1
bertopang *lean / rest on* 1
bertukar *exchange* 1
berumah *reside / have a home* 1
campur *mix* 1
cuci *wash* 1
dianggap *considered* 1
dibanting *slammed* 1
dicocok *matched* 1
dilapangkan *cleared / broadened* 1
dimabuk *intoxicated* 1
dimakan *eaten* 1
dipandang *viewed / considered* 1
gigit *bite* 1
gulung *roll* 1
jatuh *fall* 1
kehabisan *run out* 1
keluar *go out / exit* 1
kena *affected / hit* 1
lepas *detached / free* 1
lulus *pass (exam)* 1
lupa *forget* 1
main *play* 1
makan *eat* 1
mandi *bathe / shower* 1
mati *die* 1
melancarkan *launch / initiate* 1
melanjutkan *continue* 1
melicinkan *smoothen* 1
melimpahkan *bestow / transfer* 1
melirikkan *glance at* 1
meluangkan *make time* 1
melunakkan *soften* 1
meluncurkan *launch* 1
memadamkan *extinguish* 1
memancing *fish / lure* 1
memandang *look / gaze* 1
memanjatkan *offer (a prayer)* 1
memantulkan *reflect* 1
mematangkan *mature / cook* 1
mematok *set (a price)* 1
membaca *read* 1
membagikan *share / distribute* 1
membakar *burn* 1
membanting *slam / throw down* 1
membelah *split* 1
memberantas *eradicate* 1
memberitahukan *inform* 1
memberlakukan *enact / impose* 1
membilas *rinse* 1
membocorkan *leak* 1
membuang *throw away* 1
membuktikan *prove* 1
membunuh *kill* 1
memburu *chase / hunt* 1
membusungkan *puff up* 1
memekikkan *scream* 1
memerintahkan *command / order* 1
meminjam *borrow* 1
meminta *ask* 1
memompa *pump* 1
mempercepat *accelerate* 1
memperebutkan *compete for* 1
memperhatikan *pay attention* 1
mempertahankan *defend / maintain* 1
mempraktikkan *practice* 1
memulangkan *return (something)* 1
memusnahkan *destroy* 1
memutar *turn / rotate* 1
menabuh *beat (a drum)* 1
menajamkan *sharpen* 1
menambah *add* 1
menanamkan *instill* 1
menancapkan *stab / thrust* 1
menangkap *catch* 1
mencabut *pull out* 1
mencampur *mix* 1
mencantumkan *include / list* 1
mencarikan *find for* 1
mencerminkan *reflect* 1
mencolok *striking* 1
mencucurkan *pour out* 1
mencurahkan *pour (attention)* 1
mendaratkan *land* 1
mendengarkan *listen to* 1
mendirikan *establish / build* 1
menduduki *occupy* 1
menebarkan *spread* 1
meneguk *gulp* 1
menelan *swallow* 1
menempati *occupy / take place* 1
menempatkan *place* 1
menemui *meet* 1
menentukan *determine* 1
menepuk *pat / clap* 1
menerjemahkan *translate* 1
menetapkan *establish* 1
meneteskan *drop (liquid)* 1
mengadu *complain / pit* 1
mengalahkan *defeat* 1
mengantongi *pocket / gain* 1

mengarahkan *direct / guide* 1
mengasah *sharpen* 1
mengecilkan *reduce / minimize* 1
mengembalikan *return* 1
mengembangkan *develop* 1
mengendarai *ride / drive* 1
mengenyam *taste / undergo* 1
mengepalkan *clench (fist)* 1
mengerahkan *deploy / mobilize* 1
mengetuk *knock* 1
menggali *dig* 1
menggantang *measure* (in ‘gantang’) 1
menggerakkan *move* 1
menggosokkan *rub* 1
menghabiskan *spend / exhaust* 1
menghasilkan *produce / yield* 1
menghela *sigh / pull* 1
menghembuskan *blow out* 1
mengindahkan *heed / pay attention* 1
mengirimkan *send* 1
mengklaim *claim* 1
mengoleskan *apply (rub in)* 1
menguatkan *strengthen* 1
mengukir *carve* 1
mengunci *lock* 1
mengurut *massage / arrange* 1
menindak *take action on* 1
menjalin *establish (ties)* 1
menjamin *guarantee* 1
menjauhkan *distance / separate* 1
menjilat *lick* 1
menonton *watch* 1
menuju *head to / toward* 1
menukarkan *exchange* 1
menulis *write* 1
menuliskan *write down* 1
menumbuhkan *grow / develop* 1
menumpang *ride along / stay* 1
menunggu *wait* 1
menuntut *demand / sue* 1
menusuk *stab / pierce* 1
menutup *close / shut* 1
menyabung *pit (e.g. cocks)* 1
menyambung *connect* 1
menyatukan *unite / merge* 1
menyekap *confine* 1
menyemangati *encourage* 1
menyembunyikan *hide* 1
menyemprotkan *spray* 1
menyentuh *touch* 1
menyerahkan *hand over* 1
menyerukan *call out / urge* 1
menyetujui *approve* 1
menyiapkan *prepare* 1
menyibak *part / reveal* 1
menyikat *brush* 1
menyimpan *store / keep* 1
menyimpulkan *conclude* 1
menyinari *illuminate* 1
menyinggung *offend / touch on* 1
menyombongkan *boast* 1
menyumbang *contribute* 1
menyusun *arrange / compile* 1
merangkai *arrange / string* 1
merangkul *embrace* 1
merangsang *stimulate* 1
meraup *rake in / grab* 1
merintis *pioneer / initiate* 1
meruntuhkan *demolish / destroy* 1
minum *drink* 1
mulai *start / begin* 1
naik *rise / ride* 1
pandang *gaze / view* 1
pasang *install / set* 1
patah *break (bone)* 1
pulang *go home / return* 1
putus *break up / severed* 1
terbawa *carried away* 1
terbuka *open* 1
terburu *rushed* 1
tercium *smelled* 1
terganggu *disturbed* 1
terkunci *locked* 1
terpikat *attracted* 1
tersentuh *touched* 1
terserang *attacked / afflicted* 1
tertangkap *caught* 1
tertusuk *pierced* 1
tertutup *closed* 1
tidur *sleep* 1
tunjuk *point* 1
turun *go down / descend* 1
tutup *close / cover* 1

Appendix C

Sample list of the noun elements within Indonesian LVCs

C.1 List of the stative noun

gol	goal
kaki	foot / leg
manfaat	benefit
buku	book
efek	effect
solusi	solution
saksi	witness
daya tarik	appeal / attraction
sepeda	bicycle / bike
masalah	problem / issue
surat	letter / document
tangga	stairs / ladder
pertanyaan	question
hasil	result / outcome
pengaruh	influence / impact
pertimbangan	consideration
tegas	assertiveness / firmness
nafkah	livelihood / income
jawaban	answer / response
kopi	coffee
laporan	report
akses	access
sikap	attitude
televisi	television / TV
luka	wound / injury
contoh	example
nilai	value / score
karakter	character / trait
fokus	focus
pilihan	choice / option
gigi	tooth / teeth
sanksi	sanction / penalty
anak	child
nafas	breath
dendam	grudge / revenge
rekomendasi	recommendation
izin	permission / license
sukses	success
hadiah	gift / prize
prestasi	achievement
saran	suggestion / advice
keuntungan	profit / advantage
kredit	credit
bagian	part / section
uang	money
bau	smell / odor
agenda	agenda / schedule
vonis	verdict
posisi	position
jalan	road / street / way
inspirasi	inspiration
bantuan	help / assistance
resiko	risk
harga	price
berkah	blessing
ruang	space / room
kepuasan	satisfaction
korupsi	corruption
respon	response
kursi	chair / seat
sinyal	signal
pernyataan	statement / declaration
diri	self
hak suara	voting right
kesempatan	opportunity
isyarat	gesture / signal
ketersediaan	availability
tempat	place
keberatan	objection
keterampilan	skill
tanda	sign / mark
keadaan	condition / situation
santunan	compensation / aid
kehormatan	honor
jurusan	major / department
kasihan	pity / sympathy
perasaan	feeling / emotion
bekas	trace / former
demam	fever
kesaksian	testimony
keleluasaan	leeway / flexibility
hormat	respect / salute
perbedaan	difference
sesuatu	something
spanduk	banner
kemarahan	anger
golf	golf
tenda	tent
transportasi	transportation
nasihat	advice / counsel
kesepakatan	agreement
keberadaan	existence / presence
terkenal	fame / being famous
keberanian	courage
pendapatan	income / revenue
ketinggian	height / altitude
cadangan	reserve / backup
kebaikan	kindness / goodness
flu	flu
tangan	hand

C.2 Sample list of the eventive noun

keputusan	decision	mekanisme	mechanism
kesulitan	difficulty	usaha	effort / endeavor
bantuan	help / assistance	pembelian	purchase
tugas	task / duty	pertumbuhan	growth
waktu	time	perkembangan	development
informasi	information	komentar	comment
kegiatan	activity	tendangan	kick
pemeriksaan	examination / inspection	kesepakatan	agreement
dukungan	support	persiapan	preparation
kerusakan	damage	rencana	plan
peran	role	akhir	end
peluang	opportunity	proses	process
perjalanan	journey / trip	kajian	study / review
aktivitas	activity	gangguan	disturbance / interference
langkah	step / measure	kemajuan	progress
kunjungan	visit	latihan	exercise / practice
penelitian	research	kendali	control
tantangan	challenge	rapat	meeting
penyelidikan	investigation	inisiatif	initiative
perawatan	treatment / care	transaksi	transaction
pendekatan	approach	pidato	speech
tindakan	action	doa	prayer
kesempatan	opportunity	siang	afternoon / daytime
kesalahan	mistake / error	angin	wind
prioritas	priority	uang	money
apresiasi	appreciation	pendaftaran	registration
penjelasan	explanation	negosiasi	negotiation
tahu	knowledge / awareness	analisis	analysis
kerugian	loss / disadvantage	tanggapan	response / reply
pengawasan	supervision / monitoring	kuliah	lecture / college
perbuatan	act / deed	iklan	advertisement
koordinasi	coordination	pembicaraan	talk / discussion
kecelakaan	accident	suara	voice / vote
ulasan	review	olahraga	sport
perhatian	attention	tekanan	pressure
perlindungan	protection	peringatan	warning
layanan	service	penangkapan	arrest / capture
ujian	exam / test	nafas	breath
pertemuan	meeting	pertolongan	aid / rescue
operasi	operation	dorongan	push / encouragement
pekerjaan	job / work	kekurangan	shortage / lack
perbaikan	repair / improvement	penggeledahan	search (by authorities)
pengecekan	checking / verification	pemukulan	beating
upaya	effort	sejenak	moment
kehidupan	life	pembunuhan	murder / killing
pendidikan	education	promosi	promotion
sambutan	reception / welcome speech	pengobatan	treatment / medication
musik	music	pembenahan	improvement / fixing
kekalahan	defeat	percobaan	experiment / trial
serangan	attack	penyerangan	assault

Appendix D

List of LVCs with the verbs *melakukan* ‘do’, *memberikan* ‘give’, *mengambil* ‘take’, or *membawa* ‘bring’

D.1 LVCs marked by verb *melakukan* ‘do’

melakukan kajian	do a study	melakukan permainan	to play
melakukan latihan	do exercise	melakukan pelacakan	made an investigation
melakukan transaksi	do transaction	melakukan penutupan	to finish
melakukan pendaftaran	do registration	melakukan perjanjian	do promise
melakukan negosiasi	do/be.in negotiation with	melakukan lawatan	make a visit
melakukan analisis	do analysis	melakukan lompatan	make a leap
melakukan pembicaraan	do/conducting a talk	melakukan percakapan	have a conversation
melakukan olahraga	do exercise	melakukan observasi	make an observation
melakukan penangkapan	do arrest	melakukan pengrusakan	to destroy
melakukan korupsi	to commit corruption	melakukan adaptasi	inflict all.kinds.of fine-tuning
melakukan penggeledahan	make an examination	melakukan pemeliharaan	give shelter
melakukan pemukulan	to beat	melakukan syuting	get film
melakukan pembunuhan	to commit murder	melakukan penyiapan	be put up for discussion
melakukan promosi	do promotion	melakukan penyergapan	make ambushes
melakukan pembenahan	to make a revision	melakukan rekonstruksi	make extensive reconstruct
melakukan percobaan	conducts a trial	melakukan pembantaian	do murder
melakukan penyerangan	to (do) attack	melakukan klaim	claim
melakukan pengujian	do experiment	melakukan penyelaman	do a dive
melakukan perekaman	get video	melakukan penolakan	do reject
melakukan reformasi	make a reform	melakukan peletakan	stand in place
melakukan survey	to survey	melakukan pembedahan	make an operation
melakukan transfer	to transfer	melakukan musyawarah	conducting negotiations
melakukan pemesanan	put order	melakukan atraksi	proceed game
melakukan eksplorasi	perform exploration	melakukan pembebasan	to release
melakukan pembersihan	to clean	melakukan pembohongan	to lie
melakukan revisi	do revision	melakukan penuntutan	to finish
melakukan diskusi	to talk	melakukan pemindahan	do migration
melakukan orasi	do lecture	melakukan pengalihan	to move (something)
melakukan penambahan	to increase	melakukan pembatalan	to cancel
melakukan penjualan	to sell	melakukan pertunjukan	do a dance
melakukan pergantian	to be handed over	melakukan gencatan (senjata)	make a truce
melakukan penyelamatan	to rescue	melakukan telaah	do research
melakukan pencatatan	be on the list	melakukan wisata	do travel
melakukan bunuh (diri)	committing suicide	melakukan pembiusan	make someone (be) sleepy
melakukan penggerebekan	to arrest	melakukan pengorbanan	make sacrifice
melakukan perencanaan	to be on the program	melakukan pembuatan	to form
melakukan senam	do aerobics	melakukan pembuktian	to prove
melakukan audiensi	pay a state visit	melakukan penyelesaian	be about to be completed
melakukan tur	make a tour	melakukan pembacaan	to read
melakukan penyitaan	to confiscate	melakukan interogasi	do interrogation
melakukan pemanasan	be hot	melakukan pengulangan	retrospect
melakukan penggalian	do exploration	melakukan gladi	stand on stage
melakukan modifikasi	make modification	melakukan penerimaan	give acceptance
melakukan perundingan	had a negotiation	melakukan pemekaran	be in (full) bloom
melakukan pendaratan	make landing	melakukan kencan	have a date
melakukan perdagangan	do trade	melakukan pembelanjaan	do shopping
melakukan penyesuaian	make an adjustment	melakukan pendinginan	give cold
melakukan meditasi	do meditate	melakukan pembetulan	to repair
melakukan pemotretan	get photo	melakukan peledakan	do blasting

melakukan pengembaraan	to take stroll	melakukan penamparan	do hit
melakukan perbincangan	have a chat	melakukan perendaman	do submersion
melakukan ikhtiar	make an attempt	melakukan pembiasaan	make adjustment
melakukan penempatan	to occupy	melakukan pengunggahan	to put on the internet
melakukan pembesaran	do enlargement	melakukan pelengkapan	do completion
melakukan obrolan	have a chat	melakukan tamasya	go out on an excursion
melakukan penghancuran	to destroy	melakukan pelesir	take a trip
melakukan penulisan	do writing	melakukan pengambangan	do float
melakukan pembuangan	make a correct disposal	melakukan fotokopi	to do a photocopy
melakukan penghakiman	judgment on	melakukan peneguhan	be beyond doubt
melakukan pemanfaatan	make use	melakukan pesiar	to have a stroll in the park
melakukan reparasi	make a refit	melakukan pelancongan	do a journey

D.2 LVCs marked by verb *memberikan* ‘give’

memberikan bantuan	to give help	memberikan laporan	give a report
memberikan informasi	provides information	memberi kesaksian	bearing witness
memberikan manfaat	to make use of something	memberikan keleluasaan	give escape
memberi kesempatan	give an opportunity	memberi hormat	respects
memberikan apresiasi	to commend outstanding ones	memberikan sesuatu	give something
memberikan penjelasan	give explanation	memberikan pujian	give compliment
memberi tahu	give someone notice of	memberikan penghormatan	reverence
memberikan dukungan	give somebody moral support	memberikan nasihat	give a suggestion
memberikan ulasan	make comment	memberikan sentuhan	give a touch
memberikan perhatian	taking note of	memberikan ceramah	give a lecture
memberikan perlindungan	take into custody	memberikan persetujuan	give authorization
memberikan hasil	give result	memberi kabar	give news
memberikan sambutan	give a speech	memberikan instruksi	give instructions
memberikan pengaruh	to come into effect	memberikan kuliah	do/give a conference
memberikan jawaban	gives an answer	memberi perintah	gives an order
memberi contoh	gives an example	memberikan kekuatan	give strength
memberikan nilai	give a grade	memberikan hukuman	put punishment
memberikan komentar	give comments	memberikan makna	give meaning
memberi dukungan	provides support	memberikan kepercayaan	give trust
memberikan rekomendasi	give a recommendation	memberikan teguran	give reprimand
memberikan peluang	provides an opportunity	memberikan pendapat	give an opinion
memberikan izin	give a permission	memberi nasihat	give a suggestion
memberikan hadiah	give present	memberikan reaksi	give an answer
memberikan saran	give advice	memberikan tugas	give an assigment
memberi peluang	give an opportunity	memberikan pidato	give speaking
memberikan inspirasi	give inspiration	memberikan penekanan	emphasize
memberikan uang	give money	memberi alasan	gives a reason
memberikan tanggapan	give answer	memberikan arti	to give meaning
memberi peringatan	give warning	memberikan keputusan	give decision
memberi ruang	gives space	memberikan paparan	gives a presentation
memberikan kepuasan	give a satisfaction	memberikan bingkisan	give a present
memberikan pertolongan	give a hand	memberikan pertanyaan	ask a question
memberikan respon	give an answer	memberikan porsi	give passage
memberikan dorongan	give an encouragement	memberikan pengakuan	have plead
memberi peringatan	give warning	memberikan imbalan	to reward with a sum of money
memberikan sinyal	give a sign	memberi makanan	feed to satiation
memberikan (hak) suara	to cast votes	memberikan persembahan	give sacrifice
memberi isyarat	give a signal	memberikan balasan	give an answer
memberikan akses	gives access	memberikan tenggat	to enforce validity
memberikan tekanan	give a press	memberi arahan	gives direction
memberikan kebebasan	give freedom	memberikan kuasa	give the procuration
memberikan santunan	give help	memberikan kinerja	provide performance
memberikan perlawanan	give fighting	memberikan nomor	gives a number
memberikan pelajaran	give a lesson	memberikan janji	give promise

memberikan kesenangan	give pleasure	memberikan tepukan	give a slap
memberi tempat	put space or time between	memberikan keberanian	give courage
memberikan kehangatan	give heat	memberikan uluran	give help
memberikan atensi	to pay attention	memberikan paraf	to sign
memberi restu	give permission	memberikan absolusi	give an absolution
memberikan deskripsi	gives a description	memberikan jangkauan	give reach
memberikan amnesti	give amnesty	memberikan tamparan	give smack
memberikan ciuman	give a kiss	memberikan tepuk tangan	give a standing ovation
memberikan sokongan	give support	memberi istirahat	give a break
memberi suap	give bribe	memberikan penggambaran	give an explanation
memberikan kursus	give a course	memberikan bacaan	give a book
memberi komando	gives an order	memberi titah	give an instruction
memberi celah	give place to	memberikan edaran	give information
memberikan pijatan	give a massage	memberikan demonstrasi	give a demonstration
memberikan khotbah	give a sermon	memberikan keteguhan	give confidence
memberikan buku	give book	memberikan sakramen	give sacrament
memberikan pelukan	give a hug	memberikan titah	give an order
memberikan derma	give assistance	memberi perkenan	give approval
memberi ciuman	give a kiss	memberikan sun	give a kiss
memberikan minuman	give a water	memberikan tarikan	give a pull
memberikan salinan	give a copy	memberikan bisikan	give a kiss
memberikan aturan	gives a rule	memberitahukan permohonan	give a petition
memberikan berkat	give a blessing	memberikan sapuan	give a sweep
memberikan rujukan	give enthusiasm	memberikan hantaman	give a hit
memberi pertanda	give a hint	memberi keriangan	give joy
memberikan anjuran	give support	memberikan sahutan	give a response
memberikan cek	give a cheque	memberi sorakan	give cheers
memberikan konferensi	give a conference	memberikan tonjokan	give a stoke
memberikan sanjungan	make a compliment	memberi pemicu	give boost
memberi utang	lends for profit	memberikan dengusan	gave a grunt
memberikan otorisasi	give authorization	memberikan tabokan	give a slap in the face
memberikan pengesahan	give consent		
memberikan belaian	give a caress		

D.3 LVCs marked by verb *mengambil* 'take'

mengambil keputusan	take a decision	mengambil jarak	take distance
mengambil langkah	takes a step	mengambil kendali	take control
mengambil tindakan	take action	mengambil pilihan	take a choice
mengambil sikap	takes a stand	mengambil putusan	take a verdict
mengambil bagian	take part	mengambil jeda	take a rest
mengambil inisiatif	take initiative	mengambil istirahat	take a break
mengambil posisi	to take a position	mengambil catatan	take notes
mengambil resiko	take risks	mengambil simpulan	draw conclusions
mengambil kesempatan	take a chance	mengambil kesepakatan	take an agreement
mengambil jurusan (belajar)	take a major	mengambil terobosan	take the iniciative
mengambil jalan (pintas)	take shortcut		

D.4 LVCs marked by verb *membawa* 'bring'

membawa berkah	bring blessings	membawa sukacita	bring joy
membawa uang	bring money	membawa kegembiraan	bring excitement
membawa kebaikan	bring goodness	membawa kesedihan	bring sadness
membawa beban	carry a burden	membawakan barang	bring something (an object)
membawa kebahagiaan	bring happiness	membawa bawaan	bring one's belonging
membawa kedamaian	bring peace	membawa warta	bring news
membawakan tarian	perform a dance		