



ANALYSIS OF NOUN DERIVATION IN THE LAMPUNG LANGUAGE USING COMPUTATIONAL LINGUISTICS

Dewi Ratnaningsih¹, Ifan Iskandar², Ratna Dewanti³

Universitas Negeri Jakarta, Indonesia¹²³

Email: dewi_9906925003@mhs.unj.ac.id

ABSTRACT

Lampung, a regional language of Indonesia, faces declining speaker numbers, limited digital documentation, and a lack of computational studies that map its word-formation processes. This study examines noun derivation in the two major Lampung dialects, Dialect A (Api) and Dialect O (Nyo), focusing on affixation patterns, semantic categories, and dialectal variation. A dictionary-based dataset of 30,256 entries was annotated for word class through manual, rule-based, and KBBI-assisted procedures, yielding 9,259 nouns that were analyzed in Python through affix extraction, frequency analysis, and semantic categorization. The results show that 42.3% of the nouns are derived forms. The prefix *ke-*, the suffix *-an*, the nasal prefix *N-(m)*, the infix *-el-*, and the confix *ke-...-an* are the most productive noun-forming affixes, each tending toward abstract, result, agent/instrument, or state meanings. The two dialects share a largely parallel derivational system and differ mainly in affix frequency, such as the higher productivity of *pe-* in Dialect A and of the infix *-er-* in Dialect O. The study offers a systematic, data-driven description of Lampung noun morphology that supports language documentation, the development of local-language teaching materials, and language technology for revitalization.

Keywords: *affixation; computational linguistics; Lampung language; morphology; noun derivation.*

INTRODUCTION

The Lampung language, one of Indonesia's regional languages with a rich cultural heritage, faces serious preservation challenges amid globalization. Regional languages are a cultural identity inseparable from the communities that speak them (Wulandari et al., 2025), and the loss of a language means the loss of a distinctive worldview held only by its speech community (Low et al., 2022; Skirgård et al., 2023). Hameed (2022) emphasizes that a regional language that is not transmitted to the younger generation enters a phase of decline that is difficult to reverse. This challenge is even more evident because, according to Rahima (2024), many languages in the archipelago, including Lampung, are vulnerable to endangerment.

Zariquiey et al. (2022) underscore the importance of data-based language documentation as an initial step toward preservation, and Budiono et al. (2023) add that the regional languages of Sumatra require special attention because their high dialectal variation also affects language stability. Seifart et al. (2018) likewise state that language documentation must be comprehensive so that all linguistic aspects are recorded. Gomashie

and Terborg (2021) emphasize the urgency of mapping the number of speakers as an indicator of language vitality, while McIntosh et al. (2018) assert that the success of preservation is largely determined by academic support through systematic research. Studying and documenting the linguistic aspects of Lampung is therefore not merely an academic necessity but also a strategic step to safeguard the continuity of its speakers' cultural heritage.

One of the fundamental morphological aspects of Lampung is noun derivation, a process that, according to Septianingtias et al. (2024) and Amrulloh and Humairo (2020), constitutes the core of morphological productivity in a language. Noun derivation in Lampung involves attaching affixes—prefixes, infixes, suffixes, or affix combinations—to base words to form new, more specific meanings; Bunau (2022) explains that affixation is a key mechanism of lexeme formation in Austronesian languages. Affixation patterns in regional languages often reflect how a community organizes semantic categories such as agent, result, instrument, or state (Tsabita et al., 2025). Mazhitaeva et al. (2020) note that derivation reveals the structure of cultural concepts within a language, while Tariq et al. (2020) state that analyzing derivation helps identify the systematic relationship between form and meaning.

For Lampung, noun derivation patterns offer insight into how speakers make sense of the objects and entities around them, in line with the relationship between morphology and cognition (Zheng et al., 2025). Aysina (n.d.) asserts that understanding the mechanism of word formation is essential for accurate descriptive linguistic analysis. From the perspective of language documentation, derivational morphology needs to be mapped in detail because it reflects the vitality and complexity of minority languages (Kieffer et al., 2013). In the computational domain, B et al. (2018) emphasize that understanding morphological structure is the foundation for building effective natural language processing models, especially for languages with highly productive affixation. Analyzing Lampung noun derivation thus not only reveals the internal richness of the language but also provides a basis for language technology, documentation, and revitalization.

In the modern context, the computational linguistic approach offers an effective and efficient method for analyzing large-scale linguistic data (Jain et al., 2024), and Solan et al. (2005) emphasize the importance of statistical algorithms in understanding language structure. Computer-based morphological processing improves precision in identifying affixes and units of meaning (Lango et al., 2021) and allows corpora to be explored at a scale impossible to achieve manually (Pollach, 2012; Chouhan et al., 2025). Corpus-based quantitative analysis enables the objective mapping of linguistic patterns such as affix frequency (Kopotev, 2021), and statistical data can uncover distributional relationships between affixes and semantic categories (Hassan Idris Brama, 2023). Computational algorithms can also reveal latent morphological rules that are not easily identified through intuition (Ahlberg et al., 2015), and modern technology can describe morphological variation



across languages in a more structured manner (Ponti et al., 2019). Such models are particularly effective for languages with productive affixation (Szabo & Kovacs, 2020) and are important for documenting minority languages vulnerable to extinction (Majumder, 2025). A computational approach therefore enables researchers to discover consistent morphological rules and variations that manual analysis might miss, providing a data-driven foundation for describing Lampung morphology.

To date, studies specifically addressing noun derivation in Lampung remain limited. Astawan and Sadwika (2020) described derivational affixes in verbs of the Lampung Abung dialect using a generative morphology framework, contributing to the mapping of affixation in one dialect. Septianingtias et al. (2025) analyzed the lexical and phonological variation of the vocabulary item 'little finger' in Pringsewu Regency, Lampung Province, but the study was limited to a single vocabulary item and did not address the derivational system. Previous studies have thus been narrow in coverage—examining affixation in a single dialect, a single word class, or a single vocabulary item—and have not provided a comprehensive picture of noun derivation across the major Lampung dialects. Moreover, computational approaches to mapping affixation patterns have rarely been applied to Lampung. This leaves a gap in comparative, data-driven research on noun derivation that could support the documentation, teaching, and technological development of the language.

This study focuses on the two major dialects of Lampung, Dialect A (Api) and Dialect O (Nyo), which have distinct geographical distributions and linguistic characteristics. Comparing noun derivation in both dialects makes it possible to identify the shared morphological patterns of Lampung while highlighting the differences between the two variants. Accordingly, this study aims to: (1) identify the affix patterns (prefixes, infixes, suffixes, and combinations) that most frequently form derived nouns in both dialects; (2) classify nouns into semantic categories based on their morphological features and meanings, namely base nouns, infixed nouns, agent/instrument nouns, general nouns ending in -an, result/action nouns, quality/state nouns, and other derived nouns; (3) determine the relationship between affixes and the characteristics of the base words to which they attach; and (4) systematically compare the morphological rules of noun derivation in Dialect A and Dialect O. The findings are expected to enrich the documentation of Lampung, provide a basis for computational language resources for both variants, and support the teaching of Lampung as a local content subject in schools.

METHOD

This study employed a computational linguistic approach to analyze noun derivation patterns in Lampung, with a focus on comparing Dialect A and Dialect O. The research was carried out in five stages: (1) data acquisition; (2) data pre-processing; (3) identification and categorization of affixes; (4) analysis of derivation patterns; and (5) interpretation of results and their linguistic implications (Figure 1). All computational processing was performed in Python on the Google Colab platform.

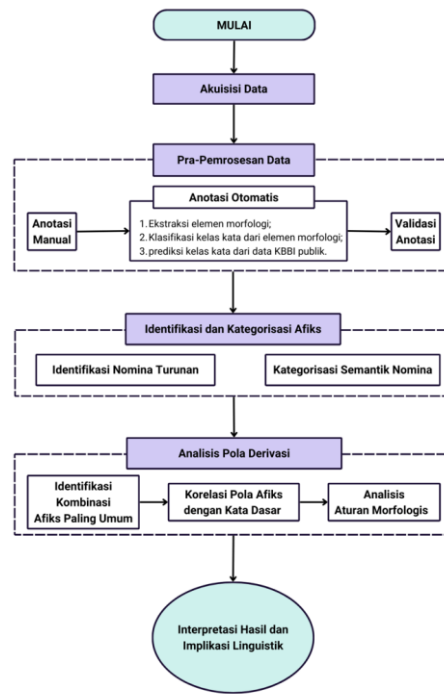


Figure 1. Stages of the Research Method

Data Sources and Acquisition

Lexical data were collected from trusted sources containing nouns and their affixes in each dialect, including digital dictionaries, linguistic documents, and field observations. Supporting data were obtained from a literature review of Lampung morphology and from interviews with native speakers and local language experts. The initial dataset consisted of 30,256 entries with eight attributes: *id*, *indonesia_kata*, *lampung_kata*, *aksara*, *dialek*, *tanggal*, *sumber_buku*, and *kelas_kata*. Because the data came from diverse sources, they were not uniform in form, particularly in spelling and dialectal variation, and were therefore normalized before analysis. In addition, word-class information (part-of-speech tags) was unavailable for most entries, so a layered annotation procedure was required before nouns could be selected for analysis.

Data Pre-processing and Word-Class Annotation

Pre-processing was carried out in structured stages, each implemented as a dedicated script whose output became the input of the next stage (Figure 2). The structure of the dataset was first validated, including the presence of the *kelas_kata* column, and new columns (*id_awalan*, *id_sisipan*, and *id_akhiran*) were added to store the affixation information needed for word-class identification.

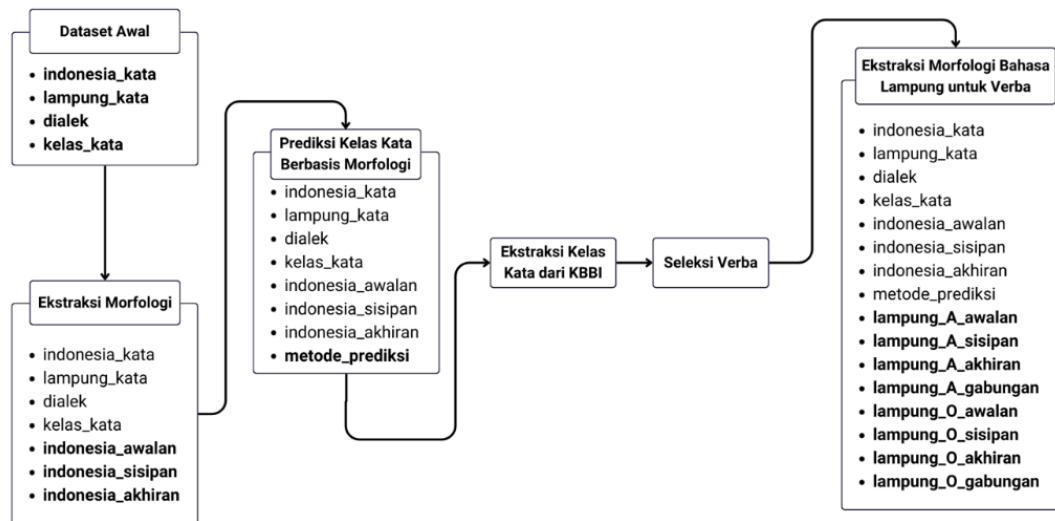


Figure 2. Stages of Data Pre-processing

Word-class annotation combined manual and computational procedures. Manual annotation was applied to a subset of the data to ensure tagging accuracy and to build reference data (a gold standard). Computational annotation was then carried out in three steps: (1) extraction of morphological elements; (2) word-class classification based on morphological elements; and (3) word-class classification based on KBBI data. Morphological extraction matched each entry against three affix inventories—prefixes (21 types), infixes (4 types), and suffixes (14 types)—with the longest affixes matched first to obtain the most specific identification and avoid overlap with shorter affixes. Word-class prediction from morphology was applied only to unannotated entries and followed a priority rule: confixes first, then single prefixes, and finally suffixes; the method used for each entry was recorded in a *metode_prediksi* column. Entries that remained unclassified were matched with their Indonesian equivalents in the KBBI using regular expressions and NLP techniques. Finally, the researchers validated every annotated entry, and the entries classified as nouns were selected for analysis.

Feature Engineering and Affix Identification

Because Lampung has a distinctive affixation system whose realizations differ between dialects, morphological features were engineered separately for each dialect. Four groups of features were created: prefix features (*lampung_A_awalan* and *lampung_O_awalan*), infix features (*lampung_A_sisipan* and *lampung_O_sisipan*), suffix features (*lampung_A_akhiran* and *lampung_O_akhiran*), and affix-combination features (*lampung_A_gabungan* and *lampung_O_gabungan*). The prefix features record the presence and type of prefixes such as *ke-*, *pe-*, *be-*, *se-*, and the nasal prefix *N-*; the infix features capture word-internal affixes such

as *-el-*, *-em-*, *-en-*, and *-er-*; the suffix features capture word-final affixes such as *-an*, *-on*, *-ni*, and *-ko*; and the combination features represent words with more than one affix, including confixes such as *ke-...-an*, *pe-...-an*, and *N-...-an*. A derivation-status variable, *is_derived*, was then constructed: a noun was classified as derived if at least one affix was detected in any of these eight feature columns.

Semantic Categorization

To classify nouns by function and derivation pattern, a semantic category function grouped each entry into seven categories (Table 1). Categories were assigned hierarchically: an affix with a more specific semantic function was given priority over a more general one, so that each noun was placed in the category most representative of its morphological characteristics.

Table 1. Semantic Categories of Noun Derivation in the Lampung Language

Noun Category	Description	Example Affix/Process
Base Noun	A noun without a derivational affix; the basic lexical form	—
Infix Noun	A noun formed through infixation	infixes <i>-el-</i> , <i>-em-</i> , <i>-en-</i>
Agent/Instrument Noun	A noun indicating the agent of an action or an instrument	prefixes <i>pe-</i> , <i>peng-</i> , <i>pen-</i> , nasal allomorph <i>N-</i>
General Noun Ending in <i>-an</i>	A noun with the productive suffix <i>-an</i> , forming a generic noun class	suffix <i>-an</i>
Result/Action Noun	A noun derived with a confix, indicating the result or abstraction of an action	confixes <i>pe-...-an</i> , <i>ke-...-an</i> , <i>per-...-an</i>
Quality/State Noun	A noun marking an abstract quality or state	prefix <i>ke-</i> , confix <i>ke-...-an</i>
Other Derived Noun	A derived noun that does not meet criteria (1)–(6) but contains an identified affix	an identified affix

Derivation Pattern Analysis and Tools

The derivation pattern analysis consisted of three stages: (1) data initialization, in which the noun dataset was loaded from a CSV file and the total number of nouns and the proportions of Dialect A and Dialect O entries were calculated; (2) analysis of affix distribution and categorization, in which each row was normalized and checked successively for prefixes, infixes, suffixes, and combinations while a frequency counter was updated; and (3) generation of the final report with sample data (Figure 3). Affix productivity was measured by occurrence frequency, and the proportion of each affix was calculated against the total number of nouns. The analysis used pandas for data manipulation and frequency counts, regular expressions for detecting affixation patterns, NLTK for basic text processing, and matplotlib and seaborn for visualization.

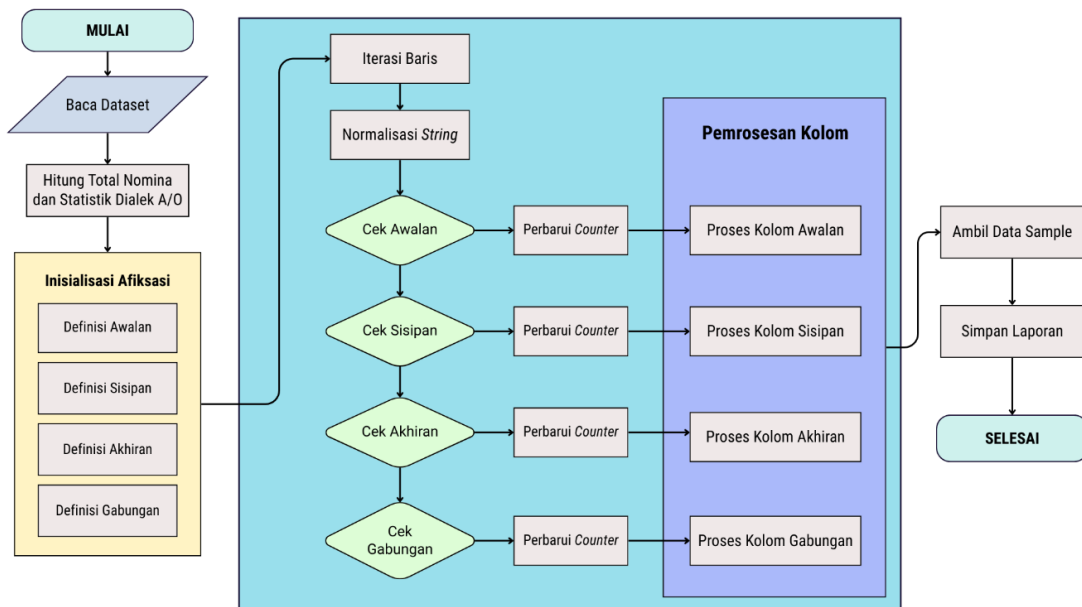


Figure 3. Stages of the Noun Derivation Analysis

RESULTS AND DISCUSSION

Corpus Characteristics

After pre-processing, 9,259 noun entries were obtained, consisting of 4,448 Dialect A nouns (48%) and 4,811 Dialect O nouns (52%). Of these, 5,343 (57.7%) were base words and 3,916 (42.3%) were derived words. Figure 4 shows the number of nouns by affixation type. Prefixation is the dominant strategy in both dialects (1,287 nouns in Dialect A and 1,321 in Dialect O, about 14% of all nouns each), followed by infixation and suffixation, while affix combinations are the least frequent. Dialect O has more infixated nouns than Dialect A (642 vs. 465; 6.93% vs. 5.02%), whereas Dialect A has more suffixed nouns (507 vs. 442; 5.48% vs. 4.77%). Forming nouns with more than one affix is therefore less productive than single affixation.

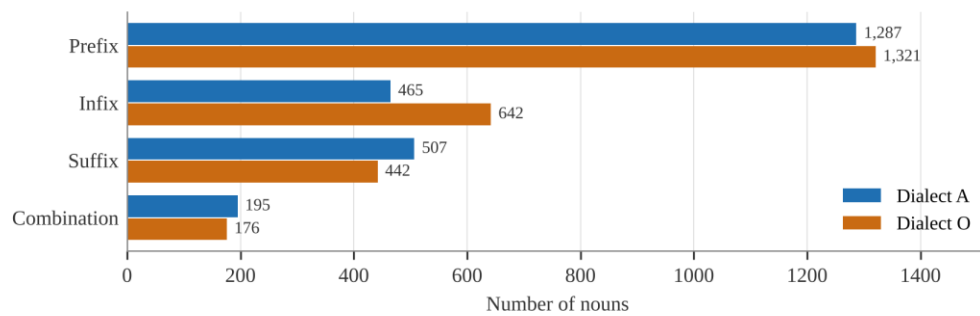


Figure 4. Number of Nouns by Affixation Type in Dialect A and Dialect O

Productivity of Affixes

Figure 5 compares the productivity of each affix type in the two dialects, and Table 2 reports the frequencies and percentages of the five most frequent affixes per type. Among the prefixes, *ke-* and *N-(m)* are the most productive in both dialects, followed by *pe-*, *se-*, *be-*, and *te-*. Their ranking differs: *pe-* ranks third in Dialect A (155) but fifth in Dialect O (119), whereas *se-* and *be-* rank higher in Dialect O. Some prefixes, such as *bu-*, *ti-*, and *pem-*, are among the fifteen most frequent only in Dialect A, whereas *per-*, *di-*, and *ber-* are among them only in Dialect O. The infix *-el-* is the most dominant infix in both dialects, but *-er-* is far more productive in Dialect O (169) than in Dialect A (17), which is the most striking difference between the dialects. The suffix *-an* dominates both dialects, while other suffixes have low frequencies, and *ke-...-an* is the most frequent affix combination, followed by *N-...-an*, *pe-...-an*, *se-...-an*, and *be-...-an*.

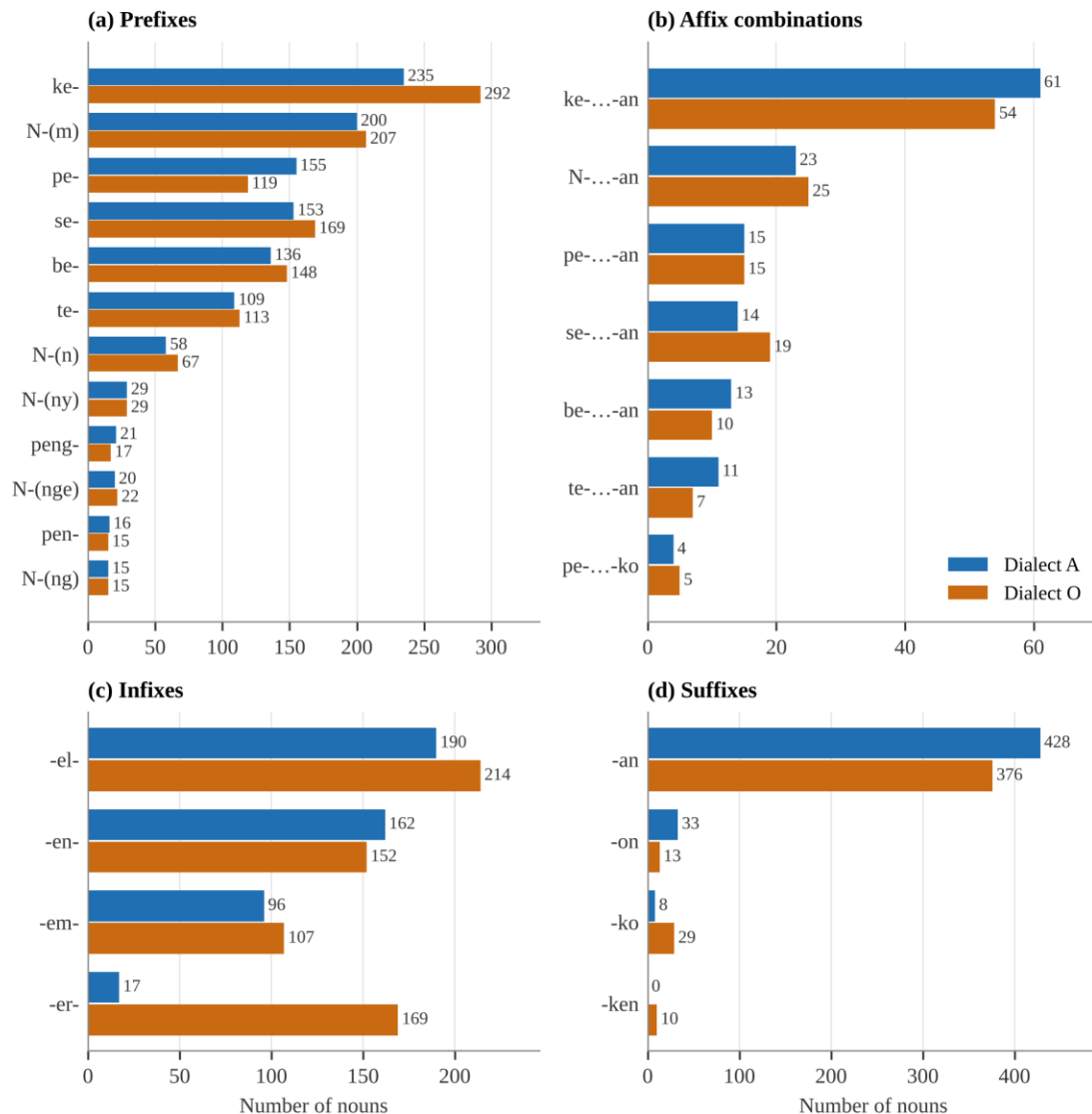


Figure 5. Productivity of Noun Affixes in Dialect A and Dialect O

Table 2. Frequency of the Most Productive Noun Affixes in Dialect A and Dialect O

Type	Affix	Dialect A n (%)	Dialect O n (%)	Notes
Prefix	ke-	235 (2.54%)	292 (3.15%)	Most productive prefix in both dialects; forms abstract nouns
	N-(m)	200 (2.16%)	207 (2.24%)	Core former of agent/instrument nouns
	pe-	155 (1.67%)	119 (1.29%)	More dominant in Dialect A
	se-	153 (1.65%)	169 (1.83%)	Unity or a single unit; stable across dialects
	be-	136 (1.47%)	148 (1.60%)	Action or state; top five in Dialect O
Infix	-el-	190 (2.05%)	214 (2.31%)	Most dominant infix in both dialects
	-en-	162 (1.75%)	152 (1.64%)	Stable distribution

Type	Affix	Dialect A n (%)	Dialect O n (%)	Notes
Suffix	-em-	96 (1.04%)	107 (1.16%)	Moderately productive
	-er-	17 (0.18%)	169 (1.83%)	Striking inter-dialect difference
	-an	428 (4.62%)	376 (4.06%)	Very dominant; forms result or general nouns
	-on	33 (0.36%)	13 (0.14%)	Low productivity
	-ni / -na	–	–	Dialect A only in the top five (19 / 17)
	-ko	8 (0.09%)	29 (0.31%)	More frequent in Dialect O
	-no / -ken	–	–	Dialect O only in the top five (14 / 10)
Combination	ke-...-an	61 (0.66%)	54 (0.58%)	Most common combination; state/abstract nouns
	N-...-an	23 (0.25%)	25 (0.27%)	Consistent across dialects
	pe-...-an	15 (0.16%)	15 (0.16%)	Abstract nouns
	se-...-an	14 (0.15%)	19 (0.21%)	Mid-level frequency
	be-...-an	13 (0.14%)	10 (0.11%)	Rare but consistent

Overall, *ke-*, *-an*, *-el-*, and *N-(m)* are the most productive noun-forming affixes in both dialects. Although most nouns are base forms without an affix (57.7%), the affixes *-an* and *N-(m)* consistently mark result and agent meanings. Double-affix patterns such as *ke-* + *-el-* and *be-* + *-el-* also occur, and the morphological patterns of the two dialects are generally similar, with slight differences in affixes such as *bu-*, *ti-*, and *per-* + *-er-*.

Correlation between Affixes and Base Words

Table 3 summarizes the semantic functions of the most common affixes and the word classes of their bases. The suffix *-an* mainly derives nouns from verbs, denoting the result of a process, an object, or a concept. The prefix *pe-* forms instrument and agent nouns, the confix *ke-...-an* forms abstract or state nouns from nouns, adjectives, and verbs, and *se-...-an* forms nouns of relation or state. In Dialect O, the nasal prefix *N-(m)* is associated with result, instrument, or activity meanings. These associations indicate that each affix selects bases of particular word classes and assigns them a relatively stable nominal meaning.

Table 3. Semantic Functions of the Main Noun-Forming Affixes

Dialect	Affix	Semantic Function	Examples (base → derived noun)	Base Class
A	-an	Result of a process	atogh ‘arrange’ → atoghan; rikin ‘count’ → rikanan	Verb
A	-an	Object or agent; concept	andol ‘rely on’ → andolan; anggep ‘consider’ → anggepan	Verb
A	-an	Abstract state or feeling	asih ‘affection’ → asihan	Adjective/noun
A	pe-	Instrument or object	kakas ‘work tool’ → pekakas; ghabut ‘household object’ → peghabut	Noun
A	pe-	Abstract (advice)	tuwah ‘counsel’ → petuwah	Noun
A	pe-	Agent or agent’s quality	gung ‘hold’ → pegung; ghassin ‘hard’ → peghassin	Verb/adjective



Dialect	Affix	Semantic Function	Examples (base → derived noun)	Base Class
A	ke-...-an	State, quality, or result	buwai 'child' → kebuwaian; pungah 'arrogant' → kepunggahan; tutung 'scorched' → ketutungan	Noun/adjective/verb
A	se-...-an	Relation or state	wari 'sibling' → sewarian; tunduk 'obedient' → setundukan; ghamok 'lively' → seghamokan	Noun/adjective
O	-an	Result of a process; quality	asah 'sharpen' → asahan; dandan 'adorn' → dandanan; amal 'devout' → amalan	Verb/adjective
O	N-(m)	Result, instrument, or activity	peppoh 'wash' → meppoh; padu 'mix' → maduken; padan 'negotiate' → madanken	Verb
O	pe-	Agent or instrument	juk 'give' → pengejuk; kajuk 'poke' → pengajuk; mutil 'pick' → pemutil	Verb

Comparison between Dialects

Table 4 compares the derivational patterns of the two dialects. Both dialects share the same main affixes and the same semantic functions for *-an*, *N-(m)*, and *se-*; the differences lie mainly in the frequency and rank of certain affixes rather than in the inventory of morphological rules.

Table 4. Comparison of Noun Derivation between Dialect A and Dialect O

Aspect	Similarities	Dialect A	Dialect O
Prefixes	ke- and N-(m) most productive; pe-, se-, be-, te- highly productive	pe- ranks higher; bu-, ti-, pem- among the top 15	se- and be- rank higher; per-, di-, ber- among the top 15
Infixes	-el- most dominant; -em- and -en- productive	-er- rare (17)	-er- highly productive (169); more infixed nouns overall
Suffixes	-an dominant, forming result/general nouns	-on, -ni, -na secondary	-ko, -no, -ken secondary
Agent/instrument nouns	N-(m), N-(n), N-(ny) are the main formers	N-(nge) and N-(ng) more prominent	More varied combinations such as N-...-an
Result/action nouns	ke-...-an and pe-...-an most consistent	Rare variants (per-...-er-...-an, ke-...-er-...-an)	Similar rare variants
Category distribution	Base nouns are the largest category; infixed and agent/instrument nouns frequent	Quality/state nouns present	Quality/state nouns absent (0)

Discussion

The findings indicate that noun derivation in Lampung relies mainly on single affixes, particularly prefixes and the suffix *-an*, whereas multi-affix combinations are less productive. This is consistent with the view that affixation is the key mechanism of lexeme formation in Austronesian languages (Bunau, 2022). The most productive affixes—*ke-*, *pe-*, *N-*, *-an*, and *ke-...-an*—have formal and functional counterparts in Indonesian, suggesting a shared derivational base, while infixes such as *-el-*, *-em-*, and *-er-*, which are no longer productive in Indonesian, remain comparatively frequent in Lampung. The association between particular affixes and semantic categories such as result, agent/instrument, and state supports

the claim that derivation reflects a systematic relationship between form and meaning (Tariq et al., 2020) and the way a community organizes its semantic categories (Tsabita et al., 2025).

At the dialect level, the two dialects share a largely parallel derivational system, differing mainly in affix frequency rather than in their inventory of morphological rules. This complements earlier research that documented Lampung variation at the lexical level (Septianingtias et al., 2024) by showing that variation also appears in the distribution of derivational affixes, most notably the infix *-er-* in Dialect O. Methodologically, the study shows that a rule-based computational pipeline can map the derivational morphology of a low-resource regional language at a scale that manual analysis cannot easily achieve (Pollach, 2012; Ponti et al., 2019). Pedagogically, the inventory of productive affixes and their meanings can serve as a basis for Lampung teaching materials and vocabulary resources for local content instruction. These interpretations should, however, be read in light of the data limitations: the corpus is dictionary-based and decontextualized, and part of it was annotated through rule-based and KBBI-assisted prediction, so the frequencies reflect lexicographic rather than spoken usage.

CONCLUSION

This study analyzed noun derivation in Dialect A and Dialect O of the Lampung language. First, the most frequent noun-forming affixes in both dialects are the suffix *-an*, the nasal prefixes *N-(m/n/ny)*, the prefixes *ke-*, *pe-*, *be-*, and *se-*, and the infixes *-el-* and *-em-*, while combinations such as *ke-...-an* and *pe-...-an* also appear consistently. Second, base nouns form the largest category in both dialects, whereas other derived nouns, infixed nouns, and agent/instrument nouns are the most frequent derivational categories; result/action and quality/state nouns are less frequent. Third, affixes correlate with the characteristics of their bases: *-an* tends to derive nouns of result, object, or concept, mostly from verbs; *N-* and *pe-* mostly form agent or instrument nouns; and *ke-...-an* forms abstract or state nouns. Fourth, the two dialects share a parallel derivational system and differ mainly in affix frequency, such as the greater prominence of *pe-* in Dialect A and of *be-* and the infix *-er-* in Dialect O, as well as the absence of quality/state nouns in the Dialect O data. This study is limited by its reliance on dictionary data and on automatic word-class prediction. Future research should use natural text corpora, validate the annotation with native speakers, and examine how these findings can be applied to Lampung language teaching materials.

BIBLIOGRAPHY

- Ahlberg, M., Forsberg, M., & Hulden, M. (2015). Paradigm classification in supervised learning of morphology. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1024–1029. <https://doi.org/10.3115/v1/N15-1107>
- Amrulloh, M. A., & Humairo, L. (2020). Morpho-phonological of loanwords in Lampung language from Arabic. *Arabiyat: Jurnal Pendidikan Bahasa Arab dan Kebahasaaraban*, 7(1), 78–97. <https://doi.org/10.15408/a.v7i1.13174>



- Astawan, ... & Sadwika, ... (2020). (Complete this reference — cited in the Introduction but missing from the original list.)
- Aysina, M. (n.d.). Word formation and structural features of the word. *EPRA International Journal of Multidisciplinary Research (IJMR)*. Retrieved November 26, 2025, from <https://eprajournals.com/IJMR/article/14844/abstract>
- B, P., K. P., S., & Kumar, M. A. (2018). A deep learning approach for Malayalam morphological analysis at character level. *Procedia Computer Science*, 132, 47–54. <https://doi.org/10.1016/j.procs.2018.05.058>
- Budiono, S., Novita, R., & Syarfina, T. (2023). Mentawai language variations in the Mentawai Islands Regency, West Sumatra Province. *Jurnal Arbitrer*, 10(1), 8–18. <https://doi.org/10.25077/ar.10.1.8-18.2023>
- Bunau, E. (2022). Morphology of the languages: Overview of the exemplifications of English, Indonesian, Iban, and Kanayatn. *Journal of English Educational Study (JEES)*, 5(1), 77–85. <https://doi.org/10.31932/jees.v5i1.1624>
- Chouhan, A., Mandour, S., & Gertz, M. (2025). ClusterChat: Multi-feature search for corpus exploration. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 4081–4085. <https://doi.org/10.1145/3726302.3730137>
- Gomashie, G. A., & Terborg, R. (2021). Nahuatl, selected vitality indicators and scales of vitality in an Indigenous language community in Mexico. *Open Linguistics*, 7(1), 166–180. <https://doi.org/10.1515/opli-2021-0010>
- Hameed, A. (2022). Mother, mother tongue, and language endangerment process: An exploratory study. *Theory and Practice in Language Studies*, 12(4), 726–735. <https://doi.org/10.17507/tpls.1204.13>
- Hassan Idris Brama, A. (2023). Analyzing the impact of sample size and population distribution on statistical test power. *International Journal of Science and Research (IJSR)*, 12(8), 2383–2390. <https://doi.org/10.21275/SR23821121418>
- Jain, D. K., Eyre, Y. G.-M., Kumar, A., Gupta, B. B., & Kotecha, K. (2024). Knowledge-based data processing for multilingual natural language analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(5), 1–16. <https://doi.org/10.1145/3583686>
- Kieffer, M. J., Biancarosa, G., & Mancilla-Martinez, J. (2013). Roles of morphological awareness in the reading comprehension of Spanish-speaking language minority learners: Exploring partial mediation by vocabulary and reading fluency. *Applied Psycholinguistics*, 34(4), 697–725. <https://doi.org/10.1017/S0142716411000920>

- Kopotev, M. V. (2021). Some thoughts on corpus and general linguistics. *Philological Class*, 26(2), 90–102. <https://doi.org/10.51762/1FK-2021-26-02-07>
- Lango, M., Žabokrtský, Z., & Ševčíková, M. (2021). Semi-automatic construction of word-formation networks. *Language Resources and Evaluation*, 55(1), 3–32. <https://doi.org/10.1007/s10579-019-09484-2>
- Low, D. S., McNeill, I., & Day, M. J. (2022). Endangered languages: A sociocognitive approach to language death, identity loss, and preservation in the age of artificial intelligence. *Sustainable Multilingualism*, 21(1), 1–25. <https://doi.org/10.2478/sm-2022-0011>
- Majumder, E. (2025). CulturalEcho: Multimodal deep learning approaches to language preservation, documentation, and community-led revitalization. *International Journal of Scientific Research in Engineering and Management*, 9(3), 1–9. <https://doi.org/10.55041/IJSREM43064>
- Mazhitaeva, S. M., Vesselinov, D., & Amanbekova, D. Y. (2020). Language representation and the structure of concepts in the modern cultural linguistics. *Bulletin of the Karaganda University. Philology Series*, 99(3), 58–64. <https://doi.org/10.31489/2020ph3/58-64>
- McIntosh, E. J., Chapman, S., Kearney, S. G., Williams, B., Althor, G., Thorn, J. P. R., Pressey, R. L., McKinnon, M. C., & Grenyer, R. (2018). Absence of evidence for the conservation outcomes of systematic conservation planning around the globe: A systematic map. *Environmental Evidence*, 7(1), 22. <https://doi.org/10.1186/s13750-018-0134-2>
- Pollach, I. (2012). Taming textual data: The contribution of corpus linguistics to computer-aided text analysis. *Organizational Research Methods*, 15(2), 263–287. <https://doi.org/10.1177/1094428111417451>
- Ponti, E. M., O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E., & Korhonen, A. (2019). Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3), 559–601. https://doi.org/10.1162/coli_a_00357
- Rahima, A. (2024). Language documentation as a revitalization effort: Regional languages are starting to extinct. *International Journal of Multidisciplinary Sciences and Arts*, 3(3), 72–76. <https://doi.org/10.47709/ijmdsa.v3i3.4123>
- Seifart, F., Evans, N., Hammarström, H., & Levinson, S. C. (2018). Language documentation twenty-five years on. *Language*, 94(4), e324–e345. <https://doi.org/10.1353/lan.2018.0070>



- Septianingtias, V., Wahya, Nur, T., & Ariyani, F. (2024). Lexical variation in the Lampung language, Indonesia. *Cogent Arts & Humanities*, 11(1), 2309740. <https://doi.org/10.1080/23311983.2024.2309740>
- Septianingtias, V., et al. (2025). (Complete this reference — cited in the Introduction but missing from the original list.)
- Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latache, J. J., Lesage, J., Weber, T., Witzlack-Makarevich, A., Passmore, S., Chira, A., Maurits, L., Dinnage, R., Dunn, M., Reesink, G., Singer, R., Bower, C., Epps, P., Hill, J., ... Gray, R. D. (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16), eadg6175. <https://doi.org/10.1126/sciadv.adg6175>
- Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, 102(33), 11629–11634. <https://doi.org/10.1073/pnas.0409746102>
- Szabo, G., & Kovacs, L. (2020). Automated learning of Hungarian morphology for inflection generation and morphological analysis. *Indonesian Journal of Electrical Engineering and Informatics (IJEI)*, 8(4), 746–756. <https://doi.org/10.11591/ijeel.v8i4.2545>
- Tariq, T. R., Rana, M. A., Sultan, B., Asif, M., Rafique, N., & Aleem, S. (2020). An analysis of derivational and inflectional morphemes. *International Journal of Linguistics*, 12(1), 83. <https://doi.org/10.5296/ijl.v12i1.16084>
- Tsabita, N. D., Amirullah, M. A., Muhdlar, H. Z. A., & Suaedi, H. (2025). Analisis penggunaan afiksasi bahasa Madura dalam buku paket Sekkar Anom 3 Bab 1 kelas IX SMP/MTs. *BLAZE: Jurnal Bahasa dan Sastra dalam Pendidikan Linguistik dan Pengembangan*, 3(1), 171–184. <https://doi.org/10.59841/blaze.v3i1.2274>
- Wulandari, A., Mulyana, D., Hadisiwi, P., & Rizal, E. (2025). Negotiating identity through language: The meaning of Javanese speech levels among teenagers in Yogyakarta, Indonesia. *International Journal of Innovative Research and Scientific Studies*, 8(2), 2022–2032. <https://doi.org/10.53894/ijirss.v8i2.5593>
- Zariquiey, R., Oncevay, A., & Vera, J. (2022). CLD² language documentation meets natural language processing for revitalising endangered languages. *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 20–30. <https://doi.org/10.18653/v1/2022.computel-1.4>
- Zheng, C., Zhao, W., Yang, Z., Guo, S., & the Alzheimer's Disease Neuroimaging Initiative. (2025). Dysfunction in the hierarchy of morphometric similarity network in Alzheimer's disease and its correlation with cognitive performance and gene expression profiles. *Psychological Medicine*, 55, e42. <https://doi.org/10.1017/S0033291725000091>