A Computational Approach to Building a Lexicon and Analyzing Morphophonemics for the Bolinao Language

Aaron Kyle M. Aguilar
Saint Louis University
2233692@slu.edu.ph

Matt Danielle U. Bravo
Saint Louis University
2233368@slu.edu.ph

Neil Angelo Q. Briones
Saint Louis University
2233376@slu.edu.ph

Milton Junsel C. Fabe
Saint Louis University
2233048@slu.edu.ph

Keanu Sonn M. Fortaleza
Saint Louis University
2233676@slu.edu.ph

Lou Diamond T. Morados
Saint Louis University
2233672@slu.edu.ph

Albert Jannsen A. Ramos
Saint Louis University
2221023@slu.edu.ph

Ruzzlee D. Salon
Saint Louis University
2216587@slu.edu.ph

ABSTRACT

The lack of digital resources for Bolinao (also called Binubolinao), one of the minority languages spoken in the Philippines province of Pangasinan, is a major obstacle in Natural Language Processing (NLP) studies. The purpose of this study is to create an NLP-ready lexicon dataset for Bolinao. The approach begins with the development of a web scrape script for the Webonary platform developed by the Summer Institute of Linguistics (SIL), from which 21,116 lexical entries were obtained. The lexicon undergoes preprocessing to enable the mapping of part-of-speech (POS) tags defined by SIL to the Universal Part-of-Speech (UPOS) tagset. The processed lexicon is then used to reverse the morphophonemic rules outlined by Persons (2025) in order to identify the root words. The final lexicon dataset size contains 20,308 Bolinao words. To ensure reliability, the lexicon was validated using secondary corpora consisting of songs, dictionaries, and stories gathered from multiple online sources. The findings contribute to the continuous effort of developing and exploring NLP tools for the Bolinao language.

Keywords

Natural language processing, computational morphology, lexical resource development, morphophonemic analysis, corpus-based validation, universal part-of-speech tagging, rule-based morphological analysis, language resource standardization

1. INTRODUCTION

1.1 Background

The digital preservation of endangered languages has become increasingly important as globalization and technological advancement threaten linguistic diversity worldwide [1]. The Philippines, hosting approximately 3% of the world's languages within just 0.2% of the Earth's geographical area, represents one of the most linguistically diverse nations globally, making it roughly 15 times more varied linguistically than the typical country [2]. However, this linguistic treasure faces unexpected challenges, with 28 Philippine languages now classified as endangered according to Ethnologue, representing a significant increase from 13 in 2016 [3]. Among these endangered languages is Bolinao, also known as Binubolinao or Bolinao Sambal, a language spoken in the Anda and Bolinao municipalities of western Pangasinan [4].

The morphological complexity of Philippine languages, including Bolinao, presents some computational challenges. These languages present intricate morphophonemic processes involving assimilation, dissimilation, vowel reduction, gemination, and complex affixation patterns that significantly change surface word forms from their underlying morphemes [5]. These morphophonemic processes are important for understanding word formation and meaning derivation, yet they remain largely undocumented in computational frameworks for endangered languages like Bolinao.

Recent research in computational morphology have showed the effectiveness of rule-based approaches in capturing morphophonemic processes for morphologically rich languages [6]. These approaches, when combined with a compilation of lexical resources, can serve as foundational tools for developing NLP applications including spell checkers, stemmers, and machine translation systems [7]. However, the development of such computational resources requires extensive linguistic documentation and systematic morphological analysis, resources that remain scarce for endangered languages like Bolinao.

1.2 Problem Statement

The scarcity of standardized digital resources for the Bolinao language presents a barrier to both linguistic preservation and technological advancement. Unlike high-resource languages with extensive digital corpora, Bolinao lacks comprehensive, machine-readable lexicons necessary for basic NLP applications [3][7]. This prevents the creation of tools including spell checkers, search engines, and machine translation systems, which are essential for language maintenance and intergenerational transmission [6].

The challenge is further supported by Bolinao's complex morphophonemic processes involving assimilation, dissimilation, vowel reduction, and gemination, which create significant surface variations from underlying root forms [5]. Existing resources use incompatible annotation schemes that prevent integration with universal NLP frameworks, while the absence of Universal Part-of-Speech (UPOS) mappings limits interoperability with established pipelines [8]. Without proper validation mechanisms, computational tools risk creating errors or failing to capture Bolinao's full linguistic complexity.

1.3 Objectives

This paper addresses the identified resource scarcity through a comprehensive approach designed to establish a computational framework for the Bolinao language while advancing methodological frameworks for endangered language digital preservation. Specifically this paper aims to:

1. Develop a cleaned, standardized digital Bolinao lexicon with UPOS tag mappings.
2. Validate the lexicon via exact and fuzzy string matching against secondary corpora.
3. Model core morphophonemic rules computationally to recover hidden roots.
4. Confirm semantic equivalence of reversed forms to ensure linguistic validity.

1.4 Contribution

This paper delivers an extensive NLP-ready lexicon for Bolinao, comprising 20,308 lemmas with Universal POS annotations, filling a digital resource gap for the language. By mapping Bolinao’s part-of-speech tags to the Universal POS tagset, the research ensures interoperability with established NLP frameworks and pre-trained models. The work introduces a replicable, rule-based morphophonemic modeling framework that captures key processes: assimilation, dissimilation, vowel reduction, and gemination, which provides the foundation for future stemming and text-processing tools. A validation methodology, combining exact and fuzzy string matching with corpus comparison, guarantees the lexicon’s accuracy and completeness. Finally, by open-sourcing all data-processing scripts and documentation, the research establishes a reproducible pipeline for endangered-language resource development that can be readily adapted to other low-resource linguistic contexts.

2. METHODOLOGY

2.1 Data Collection and Selection

We initially focused on identifying a primary corpus to serve as the foundation for the Bolinao lexicon. A survey of available digital resources for the Bolinao language was conducted that includes online dictionaries, academic papers, and collections of local literatures.

From this survey, the Bolinao dictionary of Webonary was selected as a primary corpus. We decided based on three key criteria:

(a) Comprehensiveness: It is the most substantial and well-structured lexical resource for Bolinao developed by Summer Institute of Linguistics (SIL) currently available on the internet.
(b) Scholarly Backing: The resource is directly a pillar of the linguistic research of Persons (2025). This ensures the quality and accuracy of the data.
(c) Recency: The underlying research of the Bolinao dictionary was recently revised in (June 2025) making it the current and relevant source.

2.2 Data Extraction and Preprocessing

Following the selection of the primary data source, the lexical data was extracted from the Webonary platform utilizing a custom web scraping python script. The raw data required multi-step preprocessing and cleaning such that it assures the quality and suitable for NLP tasks.

The cleaning process was led by a set of explicit rules:

(a) First, all entries were programmatically scrutinized to remove invalid characters and symbols that are not part of the Bolinao orthography.
(b) The handling of hyphens was taken into account based on the orthographic rules described by Persons [2025]. Entries with leading or trailing hyphens were removed. Internal hyphens were preserved because of the fact that this represents a glottal stop.
(c) Any entries that described a morphological process such as an entry with a derivational direction part-of-speech label (e.g., “v > ptcp” which is a verb-to-participle derivation) were removed from the final lexicon dataset.

This preprocessing pipeline was necessary in order to transform the raw web data into a clean, usable lexical resource.

2.3 Lexicon Standardization

To maximize the utilization and future applicability of the lexicon, a standardization approach was performed on the part-of-speech (POS) labels. The original dataset contained fine-grained POS tags specific to the Webonary resource. While it is detailed, these custom tags limit the lexicon's interoperability with established natural language processing (NLP) toolkits that expect a standardized tagset.

To resolve this, a mapping was created to convert the specific Webonary tags to the course-grained Universal-Part-of-Speech (UPOS) tagset, which consists of 17 universal categories (e.g., NOUN, VERB, ADJ). This process involved assigning each custom tag to its corresponding UPOS category.

The adoption of the UPOS standard is an important step transforming the lexicon into a NLP-ready resource. With this approach, the lexicon can be seamlessly integrated into such applications like training POS taggers, developing grammar checkers, or performing syntactic analysis, without the custom parsing logic.

2.4 Lexicon Validation

The lexicon validation process used a multi-tiered computational approach designed to assess the comprehensiveness and reliability of our primary Bolinao lexicon through systematic comparison with secondary corpora. The validation methodology was grounded in established corpus-based evaluation techniques.

The validation framework used a probabilistic sampling approach with fuzzy string manipulation to consider the morphological complexity and orthographic variation characteristic of Bolinao. A statistically representative sample of 300 words was randomly extracted from the compiled secondary corpus (n=2,529 unique lexical items), which ensures a 95% confidence level with a ±5% margin of error. This sample size aligns with best practices in lexical coverage studies for low-resource languages as outlined by Nurmukhamedov & Webb [2019] in their comprehensive analysis of lexical coverage and profiling methodologies.

String similarity assessment used the Levenshtein-based SequenceMatcher algorithm with a similarity threshold of 0.7, chosen specifically to capture morphological variants while minimizing false positives. This threshold was calibrated based on pilot testing consistent with approaches documented by Corris et al. [1999] for endangered language lexicography.

The validation process used three matching protocols: (1) exact string matching for direct lexical overlap, (2) fuzzy matching for orthographic and morphological variants, and (3) manual categorization of out-of-vocabulary (OOV) items to distinguish authentic lexical gaps from extraction artifacts, following the comprehensive validation framework outlined by Dunn [2024] for large-scale corpus validation.

2.5 Lexicon Coverage Analysis

The coverage analysis methodology followed established lexical profiling techniques adapted for endangered language validation, drawing on Webb's [2021] systematic review of lexical coverage research and Nurmukhamedov and Webb's [2019] comprehensive analysis of lexical profiling methodologies. Coverage metrics were calculated using the formula: Coverage = (Exact Matches + Similar Matches) / Sample Size × 100, providing both overall coverage percentages and granular analysis of match types.

Secondary corpus stratification ensured representative coverage across different text types and registers, addressing the methodological concerns raised by Mosel [2004] regarding resource selection in endangered language projects. The analysis incorporated six distinct source categories: formal dictionaries, cultural compilations, educational materials, and phrase books, totaling 765 dictionary entries and 813 phrase pairs.

The analysis used corpus similarity measures following Dunn's [2024] methodologies for validating large geographic corpora to compare lexical distributions between primary and secondary sources. This enables identification of systematic coverage gaps and validation of source representativeness. This validation approach ensures a strong assessment of lexicon quality while accounting for the specific challenges of endangered language documentation highlighted by Corris et al. [1999].

2.6 Computational Morphological Analysis

Computational morphological analysis plays a pivotal role in understanding and processing the structural components of language. By systematically examining the formation and transformation of words, it enables the identification of underlying patterns and rules that govern word construction, which is essential for applications such as natural language processing, machine translation, and linguistic research [9][10].

The significance of computational morphology lies in its ability to automate complex linguistic analysis, providing scalable and reproducible insights through techniques such as finite state transducers and machine learning models, including deep neural networks, which have shown promising results in efficiently modeling morphological processes [11]. These computational methods facilitate both parsing and generation of word forms, enhancing tasks like part-of-speech tagging, information retrieval, and text generation [12].

In this study, we apply computational techniques from Persons [1980] to reverse-engineer the finalized Bolinao lexicon to uncover root words and underlying morphological structures. This enriches our linguistic understanding and supports the development of more effective language tools. The following methodology section details the specific computational approaches and procedures used to achieve accurate morphological reconstructions.

2.6.1 Assimilation

Assimilation is defined as the phonological process where a sound alters its features to match a neighboring sound. This section focuses solely on undoing assimilated forms to recover the original root words, following the framework of Persons [1980].

Assimilation patterns were identified across Bolinao verbs, pronouns, deictic pronouns, and linkers. These recurrent shifts were formalized into rules that specify how the assimilated segments correspond to their original patterns:

Pattern 1: Verb affixes (maN-, aN-, saN-)
The /n/ in maN-/aN-/saN- changes its place of articulation to match the first consonant of the root.

(a) *man + bayo → mambayo*
(b) *man + ka + mati → mangkamati*
(c) *an + giling → manggiling*
(d) *saN- + kata’gayan → sangkata’gayan*

Pattern 2: Pronoun prefixes (koN-, ikon-)
The /N/ assimilates to the first consonant of the pronoun root.

(a) *koN- + ta → konta*
(b) *ikon + ta → ikonta*
(c) *koN- + mi → komi*

Pattern 3: Deictic pronouns (iti, isen, itaw)
Encliticization transforms the glottal + /i/ sequence into /d/.

(a) *mo + iti → modti*
(b) *mo + isen → modsen*
(c) *mo + itaw → modtaw*

Pattern 4: Linkers (a, nin)
Linking particles contract or shift to ease phonological flow.

(a) *Mangansyon ya a anak → Mangansyon yay anak*
(b) *Aripen nako nin Dios → Aripen nakon Dios*

Assimilated forms were systematically reversed by applying the rules above. Each candidate root was restored to its original base while retaining grammatical meaning.

For example:
(a) *mambayo → bayo (“to pound”)*
(b) *ikonta → ta (“we, inclusive”)*
(c) *modti → iti (“this”*)

Candidate roots generated by reversal were compared against the documented Bolinao lexicon. Only those forms with a direct lexical match were retained, reducing false positives and ensuring linguistic validity.

In cases where more than one possible root was produced, manual validation was carried out by checking whether the reversed form retained the same meaning as the assimilated word. Only forms that preserved semantic equivalence were accepted, ensuring that the reversal process did not distort the intended sense in Bolinao usage.

2.6.2 Dissimilation

Dissimilation refers to the phonological process where a sound changes to become less similar to a neighboring sound. Following the framework of Persons [1980], the methodology for undoing dissimilation was structured into four steps.

Recurrent dissimilation patterns were identified and formalized into explicit rules:

Rule 1: Infixation of -um- after bilabial stop /b/
The bilabial stop /b/ changes to the velar stop /g/ when preceded by the infix -um-.

(a) *babali → gumabali*
(b) *bagyo → gumagyo*
(c) *balasang → gumalasang*
(d) *baya → gumaya*

Rule 2: Infixation of -um- after bilabial stop /p/
The bilabial stop /p/ changes to the velar stop /k/ when preceded by the infix *-um-*.

(a) *pa’sel → kuma’sel*
(b) *padeg → kumedeg*
(c) *puso’ → kumuso’*
(d) *puti’ → kumuti’*
(e) *piklo’ → kumiklo’*

Rule 3: Completed form -inum-
In natural speech, *-inum-* often reduces to *-inm-*, *-imm-*, or *-im-* while preserving the same function. Phonological reduction simplifies *-inum-* without altering meaning, requiring recognition of alternate surface realizations.

Using the above rules, velar stops were systematically mapped back to their original bilabial stops.

For example:

(a) *gumabali → babali*
(b) *gumagyo → bagyo*
(c) *kumuti’ → puti’*

Reconstructed roots were compared with the documented Bolinao lexicon. Only forms that directly matched lexical entries were retained as valid roots, ensuring accuracy and avoiding false positives.

Finally, manual validation checked whether the reversed forms preserved the same meaning as their dissimilated counterparts. Forms that maintained semantic equivalence were accepted, while mismatches were excluded.

2.6.3 Vowel Reduction

Vowel reduction is a morphophonemic process observed in many languages, where vowels are shortened, weakened, or deleted in certain environments. In Bolinao, this process takes place when affixes are attached, often leading to the loss or alteration of medial and final vowels in the root. This section focuses on reversing those reductions in order to recover the original root words, following the descriptions of Persons [1980].

Vowel reduction patterns were identified in Bolinao across verbs and nouns. These recurrent shifts were formalized into rules that specify how reduced segments correspond to their original forms:

Form 1: Medial Vowel Deletion
The medial vowels /a/ and /e/ may be deleted when a root word combines with affixes.

(a) *ma - + kasaw → maksaw*
(b) *galat + ed → galten*
(c) *na + busoy → nabsoy*

Form 2: Final Vowel Reduction Before Suffixes
The final vowel /i/ becomes /y/, and /o/ becomes /w/, when followed by the suffixes -an or -en.

(a) *irgo + en → irgwen*
(b) *kadiri + an → kadiryan*

Form 3: Undoing Vowel Reduction
Reduced forms were systematically reversed by applying the rules above. Each was restored to its original base while retaining grammatical meaning.

Candidate roots generated by reversal were then compared against the documented Bolinao lexicon. Only those forms with a direct lexical match were retained to reduce false positives and ensure linguistic validity.

For example:
(a) *maksaw → kasaw (“to be strong”)*
(b) *galten → galat (“anger”)*
(c) *nabsoy → busoy (“satiated, full”)*
(d) *irgwen → irgo (“to climb”)*
(e) *kadiryan → kadiri (“shame, embarrassment”)*

2.6.4 Assimilation and Reduction

Assimilation and reduction represent complex morphophonemic processes in Bolinao, where prefixes ending in /ng/ (maNg-, naNg-, paNg-) undergo systematic sound changes when attached to roots beginning with specific consonants. This process involves two simultaneous operations: the /ng/ assimilates to match the point of articulation of the following consonant, and the initial consonant of the root is deleted (reduced). The computational rules implemented follow these rules:

Rule 1: For stops /p, b/:

(a) *maNg- + /p,b/ → ma- + m + Ø (root minus initial consonant or apostrophe)*
(b) *naNg- + /p,b/ → nam- + Ø*
(c) *paNg- + /p,b/ → pam- + Ø*

Rule 2: For stops /t, d, s/:

(a) *maNg- + /t,d,s/ → man- + Ø*
(b) *naNg- + /t,d,s/ → nan- + Ø*
(c) *paNg- + /t,d,s/ → pan- + Ø*

Rule 3: For velars /k, g/ and glottal /'/:

(a) *maNg- + /k,g,'/ → mang- + Ø*
(b) *naNg- + /k,g,'/ → nang- + Ø*
(c) *paNg- + /k,g,'/ → pang- + Ø*

Rule 4: For sonorants /y, w, l, r, m, n, ng/ and vowels /a, e, i, o, u/:

(a) *No reduction occurs; /ng/ remains unchanged*

Our computational analysis processed 865 candidate words from the Bolinao lexicon that exhibited these prefix patterns. The algorithm systematically:

1. Identified words beginning with assimilated prefixes (mam-, man-, mang-, nam-, nan-, nang-, pam-, pan-, pang-)
2. Applied reverse morphophonemic rules to reconstruct potential root forms
3. Generated multiple root candidates where phonological ambiguity existed
4. Cross-validated predictions against the lexicon database

2.6.5 Pronoun Changes

The pronouns in the Bolinao language frequently surface with prefixes or even altered segments that usually obscure their underlying morphemes [5]. These alternations occur most frequently in interrogative and demonstrative pronouns, which take on forms that differ from their root structures, which would require explicit computational rules in order to recover their canonical stems.

Rule 1: Interrogative Prefix-Stripping (*an-*/*ans-*)
Interrogatives are usually marked with the prefixes an- or ans-. These prefixes are stripped in order to recover the root words:

(a) ansai → sai
(b) ansara → sara
(c) ansi → si
(d) ansaray → saray
(e) ansay → say

Rule 2: Demonstrative Normalization (*moy-*/*mod-*)
Demonstratives frequently surface with the prefixes moy- or mod-, or even undergo substitution due to linker alternations. These forms are mapped to their canonical words through explicit dictionary normalization:

(a) moyti or *modti → iti*
(b) moin → in
(c) modsen → isen
(d) moytaw → taw
(e) modtaw → itaw

Rule 3: Indefinite and Expletive Retainment
Both indefinites and expletives retain their surface forms as their roots. For the indefinites, reduplication (e.g., *inin*) shows indefiniteness and is semantically essential, not an alternation. Undoing reduplication would remove grammatical meaning, therefore, the words are preserved. Expletives on the other hand also show no alternations, therefore, the forms are also preserved.

Rule 4: Possessive Exclusion
Possessive pronouns that were captured (e.g., *nin*, *ri*, *sa*, *alas*) acted more like particles or borrowed words rather than core pronouns. And since they don't undergo alternations and already exist in canonical form, they are excluded from the normalization.

2.6.6 No. of Mass Indication

Number or mass indication is defined as the morphological process in which a word undergoes reduplication, infixation, or gemination to signal plurality or collectivity. This section focuses solely on detecting such marked forms and recovering their original root words, following the framework of Persons [1980].

Rule 1: Reduplication
A reduplicated consonant–vowel (CV) sequence at the beginning of a word is dropped to recover the root.

(a) *bubato → bato ("group of stones" → "stone")*
(b) *bubaboy → baboy ("group of pigs" → "pig")*
(c) *uanak → anak ("children" → "child")*

Rule 2: Gemination
When consonants or vowels are doubled, they are simplified to a single sound. In cases of initial gemination, redundant leading vowels or consonants are also removed.

(a) *lulalaki → lalaki* ("men" → "man")
(b) *bubbiyi → babayi* ("women" → "woman")
(c) *aanakan → anakan* ("offspring" → "child")

Rule 3: Infixation
The infixes *-u-* or *-aw-* are removed or replaced with the base form of the root.

(a) *bawbato → bato* ("stones" → "stone")

Reconstructed roots were systematically compared with the documented Bolinao lexicon. Only forms that matched established lexical entries were retained as valid roots, ensuring both accuracy and the avoidance of false positives.

After lexicon filtering, special cases such as full reduplication (*lolo*, *bibi*) were excluded, as these represent distinct lexical items rather than mass indicators. Manual review was also applied in ambiguous cases to confirm the correctness of recovered roots.

2.6.7 Gemination and Reduction

The analysis of gemination and reduction in Bolinao was carried out in two phases: an initial application across the entire lexicon dataset, and a refined application restricted to confirmed root candidates. Both stages relied on computational rules designed to capture morphophonemic processes.

Gemination was identified where consonants appeared doubled in the surface form but trace back to a single consonant in the root (e.g., sallamat → salamat).

Reduction was defined as the deletion or simplification of consonant clusters or repeated vowels in the root form, often reflected in shorter surface realizations (e.g., biyát → byat).

The rules were implemented through string-matching operations that compared word forms against a lexicon of validated roots. Candidate roots were generated by either removing doubled consonants (gemination) or restoring deleted vowels/consonants (reduction).

We applied the gemination and reduction rules to the whole dataset in the initial process. The results produced many candidates, many of which were false positives, since the system also attempted reconstruction for words with no valid gemination or reduction.

We have identified recurrent patterns of gemination and reduction in Bolinao and encoded them into explicit computational rules. These rules approximate the underlying root forms by collapsing doubled consonants or restoring vowels in consonant clusters. Unlike manual approaches, validation here relied solely on automated comparison against the documented Bolinao lexicon, without additional speaker-informed checking.

Rule 1: Consonant Gemination (CC → C)
Words containing doubled consonants were reversed by collapsing the geminate sequence into a single consonant.

(a) *mammuno → mamuno*

(b) *binnugtong → binugtong*

(c) *maggapo → magapo*

Rule 2: Vowel Reduction in Consonant Clusters (CVC → CC)
Clusters due to vowel deletion were expanded by restoring a predictable vowel, typically a or i, based on attested patterns.

(a) *dakpen → dakapen*

(b) *blangen → bilangen*

(c) *gusgusen → gusagusen*

Rule 3: Affixation-Induced Clusters (man-, pan-, san-)
Clusters introduced by common prefixes could mimic reduction. Reversal reconstructed the complete prefix sequence before isolating the root.

(a) *mannalay → manalay*

(b) *pannabunen → panabunan*

Rule 4: Apostrophe Handling
Apostrophes marking glottal stops or morpheme boundaries were normalized to avoid false detection of gemination or reduction.

(a) *a'lon → alon*

(b) *man'ipambayo → manipambayo → bayo*

Rule 5: Reduplication and False Doubling
Words containing apparent doubling due to reduplication or expressive extension were retained unless the reconstructed form had a direct lexical match.

(a) *sibibbyay → sibibiyay*

All reconstructed candidates were automatically cross-checked with the Bolinao lexicon. Only forms with a valid lexical match were retained as root candidates, while unmatched forms were flagged as potential edge cases. For example:

(a) *magte' → magate' → gate*

(b) *pmbayo → pambayo → bayo*

(c) *sallamat → salamat → salamat*

This process allowed the systematic identification of gemination and reduction patterns while reducing false positives. However, unresolved ambiguities, such as multiple possible vowel insertions, were carried forward as edge cases for further analysis.

2.7 Methodology Overview

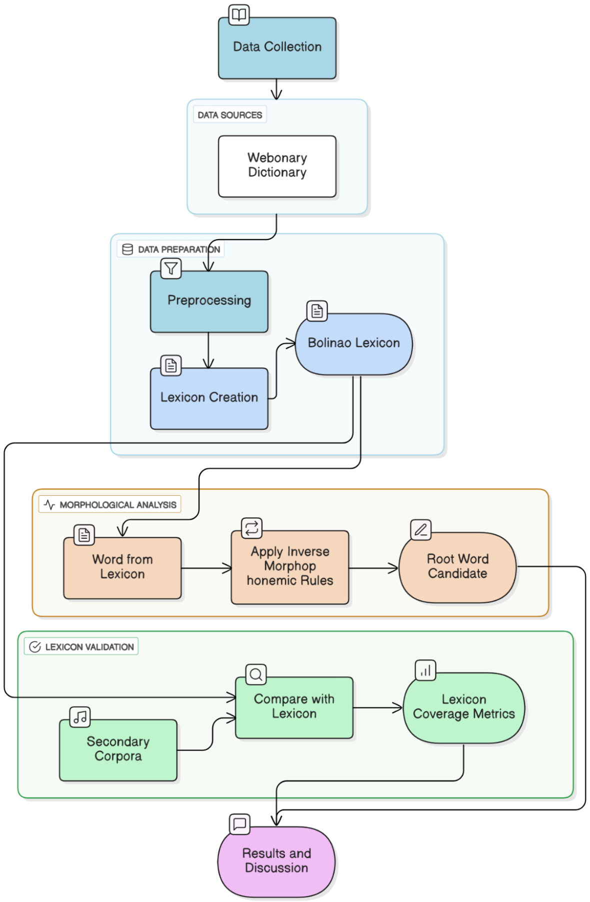

Figure 1: Methodology Workflow

In summary, the methodology integrates both computational linguistics and corpus-based validation to construct and analyze a Bolinao lexicon. Data collection starts with a primary source (Webonary Dictionary), followed by preprocessing and structured lexicon creation. Morphophonemic analysis is applied to extract root word candidates using inverse rules. For validation, the created lexicon is systematically compared against secondary textual corpora, with lexicon coverage and performance measured quantitatively. The workflow culminates in an analysis of results, facilitating further discussion and interpretation of findings.

3. RESULTS AND DISCUSSIONS

3.1 The Bolinao Lexicon: A Quantitative Overview

The data collection and preprocessing resulted in a comprehensive electronic lexicon for the Bolinao Language. The lexicon comprises a total of 20,308 entries. A validation script was executed to check for duplicate word forms. If a word form appears with different parts of speech, this corresponds to a multiple lemma entry. The result shows that the lexicon consists of 20,308 unique lemmas, indicating a clean, one-to-one mapping between a word and its definition in this resource.

To standardize the lexicon for interoperability with global NLP frameworks, each entry was mapped to a Universal-Part-of-Speech (UPOS) tag. The frequency distribution of these tags highlights the layout of the dataset. As it is common in natural language, verbs and nouns form the largest categories. The distribution is detailed in Table #, and a visual representation is provided in Figure #.

Table 1: Frequency Distribution of Universal POS Tags

UPOS Tag	Frequency	Percentage
VERB	11072	54.5231%
NOUN	7577	37.3123%
ADJ	1064	5.2396%
NUM	157	0.7731%
ADV	142	0.6993%
PART	67	0.3299%
AUX	62	0.3053%
PROPN	54	0.2659%
INTJ	36	0.1773%
PRON	32	0.1576%
ADP	26	0.1280%
CCON	9	0.0443%
SCONJ	7	0.0345%
DET	2	0.0098%
Total	20,308	100%

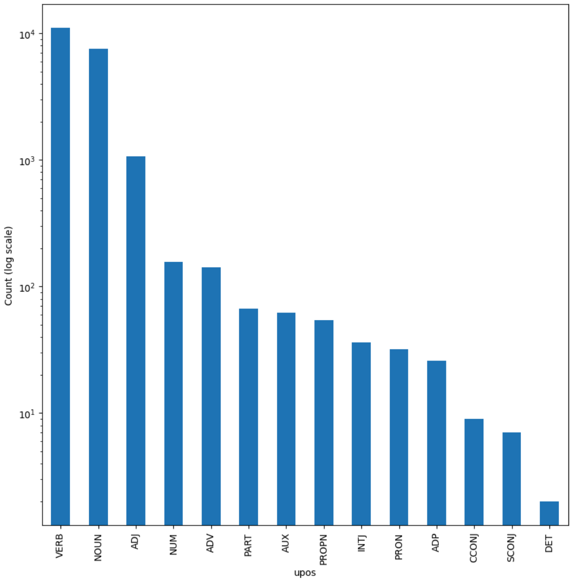

Figure 2: Distribution of UPOS tags in the Bolinao Lexicon (y-axis in logarithmic scale)

3.2 Lexicon Coverage and Error Analysis

The validation analysis revealed substantial lexical coverage with significant insights into the morphological and orthographic characteristics of Bolinao. From the 300-word random sample, 73.0% achieved coverage through either exact or similar matching, indicating strong foundational representation in our primary lexicon, consistent with coverage rates reported by Webb [2021] for well-established lexical resources.

Table 2: Quantitative Coverage Results

Metric	Count	Percentage
Total Sample Size	300	100.0%
Total Coverage	219	73.0%
Exact Matches	36	12.0%
Similar Matches (Variants)	183	61.0%
Out-of-Vocabulary (OOV)	81	27.0%

The low exact match rate of 12.0% contrasts sharply with the high variant detection rate of 61.0%, showing a substantial morphological and orthographic variation within the language. This pattern aligns with Webb's [2021] findings on lexical profiling in morphologically rich languages, where variant recognition significantly impacts coverage assessment, and reflects the challenges identified by Mosel [2004] regarding standardization issues in endangered language lexicography.

Systematic analysis of the 81 OOV items showed six distinct error categories, providing insights into validation pipeline effectiveness and addressing the methodological concerns raised by Dunn (2024) regarding automated corpus validation:

Category 1: English Contamination (28.6%)
The largest error category consisted of English words (commodities, termite, cultivate, translatable) extracted from mixed-language source materials. This finding highlights the challenge regarding automated corpus creation in multilingual contexts and underscores the importance of strong language identification protocols.

Category 2: Authentic Bolinao Terms (20.4%)
Genuine Bolinao words absent from the primary lexicon (andukéy, ugál, dásig, ansisímot, bañggíra) represent true lexical gaps requiring incorporation. These findings validate the emphasis on comprehensive source consultation for endangered language dictionaries and demonstrate the value of multi-source validation approaches.

Category 3: Morphological Variants (14.3%)
Systematic prefix and suffix variations (nalingwanan vs. kalingwanan, mampuuran vs. mapuuran) demonstrated productive morphological processes inadequately captured in the base lexicon. This aligns with the observations on morphological complexity in endangered language lexical resource creation and highlights the importance of morphological analysis in validation pipelines.

Category 4: Orthographic Variants (12.2%)
Spelling inconsistencies (mag´in vs. mag'in, bayáni vs. bayani) revealed standardization challenges common in languages with limited written traditions, consistent with the orthographic issues discussed in endangered language dictionary projects.

Category 5: Technical/Metadata Terms (14.3%)
Document artifacts represented processing noise rather than vocabulary gaps, indicating the need for improved filtering mechanisms as recommended for large-scale corpus validation.

Category 6: OCR Processing Errors (10.2%)
Malformed entries reflected optical character recognition limitations in historical document processing, highlighting technological challenges in endangered language documentation.

3.3 Analysis of Morphonemic Rules

In this section, we present the results of our study and provide insights from the reversal of the finalized Bolinao lexicon to root words. The analysis covers all examined morphophonemic processes, including assimilation, dissimilation, vowel reduction, combined assimilation and reduction, pronoun changes, number or mass indication, and gemination with reduction. By examining these processes, we can identify patterns, evaluate the productivity of rules, and assess how well semantic meaning is preserved across derivations.

3.3.1 Assimilation

Table 3: Assimilation Results

Description	Count	Percentage
Total root words identified	86	100.0%
Semantically equivalent forms	59	68.6%
Non-equivalent forms	26	31.4%

On the final root words derived through assimilation, a total of 86 root words were identified. Many of these aligned with valid root forms and preserved their meanings, showing that assimilation can function as a straightforward phonological process. However, not all forms behaved this way. Some words that appeared assimilated were already independent root words in Bolinao. For instance, *anti* is itself a root, and reversing it to *ti* produces a form that does not exist or carry the same meaning. These cases explain why not all recovered forms were semantically equivalent.

Table 4: Examples of Correct Assimilation

Assimilated Form	Root Word	Meaning of Assimilated Form	Meaning of Root Word	Type of Assimilation
mangumpisal	kumpisal	To confess (act of confession to a person or priest).	Confession of a wrong done.	maN- prefix
ansarayti	sarayti	"What are these?" (plural, near the speaker).	These specific events, persons, or things near the speaker.	aN- prefix
mammigulo	migulo	One who causes conflict; to scuffle or create commotion.	To scuffle or create commotion.	maN- prefix
sankutsara'	kutsara'	One spoon of something.	Spoon.	saN- prefix
sansalop	salop	One 3-liter can or nine condensed milk cans of something.	A standard dry measure (~3 liters).	saN- prefix

The examples from Table 4 demonstrate cases where assimilation rules in Bolinao operate without distorting the root word's meaning. In each case, the assimilated form retains the semantic core of the root while adjusting phonologically to fit affixation patterns. For instance, the use of *maN*- and *saN*- prefixes simply alters the initial consonant of the root to match the place of articulation but leaves the word's sense intact. This shows that assimilation in Bolinao is largely predictable and does not hinder intelligibility when applied correctly.

Table 5: Examples of Incorrect Assimilation

Assimilated Form	Root Word	Meaning of Assimilated Form	Meaning of Root Word	Type of Assimilation
anti	ti	Aunt.	This specific (deictic pronoun).	aN- prefix (misidentified)

santan	tan	Ixora shrub, with varieties identified by flower color (white, red, pink).	And (linker connecting elements).	saN- prefix (misidentified)
mangaso	kaso	To hunt using a dog.	A court case or legal matter.	maN- prefix (misidentified
anem’em	em’em	Warmness equal to body temperature.	To mull over hidden negative feelings.	aN- prefix (misidentified
anina	ina	Expression of extreme wonder, apprehension, or frustration.	Mother.	aN- prefix (misidentified

Table 5 illustrates instances where assimilation produced forms that diverge semantically from their supposed roots. In these cases, the reversal process failed because the assimilated words are not truly derived from the roots suggested by the rules. For example, *anti* is already a root word meaning “aunt,” and reversing it to *ti* generates a non-existent form. Similarly, *santan* (a plant name) and *mangaso* (to hunt) show mismatched meanings when linked back to *tan* and *kaso*. These cases highlight the limits of purely rule-based reversal: not all surface forms that appear assimilated are actually the result of assimilation.

Overall, the findings show that assimilation in Bolinao is a productive and reliable process. In most cases, it successfully adjusts the sound of a root word to fit with a prefix while keeping the meaning intact, as seen in forms like *mangumpisal* and *sansalop*. This shows that assimilation plays an important role in shaping words without distorting their meaning. While a smaller portion of cases did not align such as *anti* and *santan* these exceptions only highlight that not every apparent form comes from assimilation. Taken together, the results affirm that assimilation accounts for the majority of valid root formations in Bolinao, making it a key process in the language’s word structure.

3.3.2 Dissimilation

Table 6: Dissimilation Results

Description	Count	Percentage
Total root words identified	74	100.0%
Semantically equivalent forms	57	77.0%
Non-equivalent forms	17	23.0%

On the final root words derived through dissimilation, a total of 74 root words were identified. A majority of these 57 words retained semantic equivalence with their original forms, demonstrating that dissimilation can be an effective strategy for uncovering underlying roots in Bolinao. However, some forms did not preserve meaning. In certain cases, the process produced forms that are either unattested or function as independent words with different meanings. For example, a root that initially underwent consonant differentiation might yield a form that does not exist in the lexicon, highlighting why some dissimilated forms fail to be semantically equivalent.

Table 7: Examples of Correct Dissimilation

Dissimulated Form	Root Word	Meaning of Dissimulated Form	Meaning of Root Word	Type of Dissimilation
gumagay	bagay	For something to become appropriate for use or purpose	To fit together as being appropriate to each other	Rule 1: /b/ → /g/ after -um-
gumitil	bitil	To become hungry	A famine, a time without food	Rule 1: /b/ → /g/ after -um-
gumulong	bulong	To be budding and leafing out	A leaf or a portion of something similar to a leaf	Rule 1: /b/ → /g/ after -um-
kuma'sel	pa'sel	To be doing something in a moody manner	The characteristic of being temperamenta, especially in children or older people; sulking, moody, or bad-tempered	Rule 2: /p/ → /k/ after -um-
kumalet	palet	To become thick	The viscosity or thickness of a liquid	Rule 2: /p/ → /k/ after -um-

The examples from Table 7 demonstrate cases where dissimilation rules in Bolinao operate without altering the root word’s meaning. Each dissimulated form preserves the semantic core of the root while adjusting phonologically according to the infixation rules. For instance, the infix -um- triggers a predictable change of /b/ → /g/ or /p/ → /k/ depending on the initial consonant, as seen in forms like gumagay from bagay and kuma'sel from pa'sel. These changes follow systematic phonological patterns that do not obscure the meaning of the original root. This shows that dissimilation in Bolinao is a regular and intelligible process when applied correctly, allowing speakers to form derived words while maintaining semantic transparency.

Table 8: Examples of Incorrect Dissimilation

Dissimulated Form	Root Word	Meaning of Dissimulated Form	Meaning of Root Word	Type of Dissimilation
gumalet	balet	To persist on something over a long period of time	Heat rash, skin eruptions from heat	Rule 2 misapplied (/p/ → /k/)
gumaya	baya	To smolder	Glowing embers, burning charcoal	Rule 1 misapplied (/b/ → /g/)
kumusta	pusta	A greeting: Hello, How are you	Stakes in a bet	Rule 2 misapplied (/p/ → /k/)

piman	pinuman	To be in a pitiful condition	To drink together	Rule 2 misapplied (/p/ → /k/)
gumwa'	bwa'	To do or make something	The ball-like growth of a new plant inside a dry coconut	Rule 1 misapplied (/b/ → /g/)

Table 8 illustrates instances where dissimilation produced forms that diverge semantically from their supposed roots. In these cases, the dissimilation process fails because the rules are incorrectly applied or the words are not genuinely derived from the suggested roots. For example, *gumalet* from *balet* and *gumaya* from *baya* produce forms whose meanings no longer align with the root, while *kumusta* and *piman* create words that are semantically unrelated to their supposed bases. These examples highlight the limits of purely rule-based dissimilation: not every surface form that appears to follow the phonological patterns is actually a valid derivation, emphasizing the need to verify semantic equivalence alongside phonological correctness.

Overall, the findings show that dissimilation in Bolinao is a systematic and generally reliable process. In most cases, it successfully alters the initial consonant of a root word according to predictable phonological rules such as /b/ → /g/ and /p/ → /k/ after the infix -um while preserving the meaning, as seen in forms like *gumagay* and *kuma'sel*. This demonstrates that dissimilation plays an important role in forming derived words without distorting their semantic core. While a smaller portion of cases did not align, such as *gumalet* and *gumaya*, these exceptions highlight that not every surface form is a valid outcome of dissimilation. Taken together, the results affirm that dissimilation accounts for a substantial portion of valid root derivations in Bolinao, making it a key mechanism in the language's word structure.

3.3.3 Vowel Reduction

Table 9: Vowel Reduction Results

Description	Count	Percentage
Total word-root pairs generated	77	100.0%
Pairs with formal and semantic alignment	9	11.7%
Pairs weakly related or semantically unrelated	68	88.3%

From a total of 77 word-root pairs generated through the vowel reduction process, only 9 pairs (11.7%) exhibited both formal and semantic alignment between the reduced word and its candidate root. The majority of pairs (88.3%) were either weakly related or semantically unrelated, despite being formally valid entries. Part-of-speech tags matched in most cases, indicating that vowel reduction tends to preserve grammatical category even when the meaning of the word diverges from its root. These results suggest that while vowel reduction can reveal underlying morphological patterns, it is less reliable than dissimilation or assimilation in maintaining semantic transparency:

Table 10. Examples of Correct Vowel Reduction

Reduced Form	Original Form	Meaning of Reduced Form	Meaning of Original Form	Type of Vowel Reduction
pable'	pabale'	To lend or allow something to be used without payment for a period of time	To have or allow a response to something	Form 1: Medial Vowel Deletion
makaarwas	makaara was	To be able to come up out onto a surface from within something such as water or a boat	To come up out from within something such as water or a boat onto the surface	Form 1: Medial Vowel Deletion
makadpa'	makadap a'	To be able to land or alight on	To fall into a lying position with one's face down on the ground	Form 1: Medial Vowel Deletion
dayday	dayaday	To spread something out	To be laid out or hung out to dry	Form 1: Medial Vowel Deletion
idpa'	idapa'	To drop something flat on the ground	To make fall face down	Form 1: Medial Vowel Deletion

The examples from Table 10 demonstrate cases where vowel reduction in Bolinao operates without altering the meaning of the root word. Each reduced form preserves the semantic core of the original while adjusting phonologically according to predictable vowel deletion or modification patterns. For instance, medial vowels /a/ and /e/ are systematically deleted in forms like pable' from pabale' and *makaarwas* from *makaarawas*, while the final vowel changes in forms like *dayday* from *dayaday* maintain grammatical function. These changes follow consistent rules that do not obscure the meaning of the root, showing that vowel reduction is a regular and intelligible process in Bolinao when applied correctly, allowing speakers to produce concise forms while retaining semantic transparency.

Table 11: Examples of Incorrect Vowel Reduction

Reduced Form	Original Form	Meaning of Reduced Form	Meaning of Original Form	Type of Vowel Reduction
agwa	agawa	Perfume water	Diligence combined with promptness	Form 1 misapplied: Medial vowel deletion
karetket	kareteket	A crease or ridge caused by shrinkage due to burning or drying	Wind vane, an instrument used to determine the direction of the wind	Form 1 misapplied: Medial vowel deletion
bakbak	bakabak	A fight between dogs or pigs that are biting	A dry wood termite that produces frass or pellets	Form 1 misapplied: Medial vowel deletion
sabtan	sabatan	To solve or identify the answer to a riddle or a puzzle	To cut something specific with a curved-tip bolo	Form 1 misapplied: Medial vowel deletion
trak	terak	A truck	A burst, an explosion or eruption from something	Form 2 misapplied: Final vowel reduction

Table 11 illustrates instances where vowel reduction produces forms that diverge semantically from their supposed roots. In these cases, the reduction process fails because the shortened words are not truly derived from the roots suggested by the rules. For example, *agwa* from *agawa* and *karetket* from *kareteket* result in meanings that no longer align with the original forms. Similarly, *bakbak*, *sabtan*, and *trak* create words whose senses are mismatched with their supposed bases. These cases highlight the limits of purely rule-based vowel reduction: not every surface form that appears reduced reflects a valid morphological derivation, emphasizing the need to consider semantic as well as phonological criteria.

Overall, the reversal of vowel reduction produced only a very limited number of valid results. Out of 77 word-root pairs generated, only 9 showed any meaningful connection, and even those were not always consistent in terms of semantic alignment. The majority of recovered roots were either unrelated in meaning or produced false matches, highlighting the limitations of applying vowel reduction rules in isolation. This outcome reflects the fact that the rules provided by Persons [1980] are too restricted to fully capture the complexity of vowel reduction in Bolinao. Additional linguistic data, contextual information, or broader sets of morphophonemic rules are needed to improve the accuracy and meaningfulness of root word recovery.

3.3.4 Assimilation and Reduction

Table 12: Assimilation and Reduction Results

Description	Count	Percentage
Total root words identified	865	100.0%
Semantically equivalent forms	789	90.52%
Non-equivalent forms	76	9.48%

From the assimilation and reduction process, 865 root words were identified, of which 789 (90.52%) were semantically equivalent to their root forms. The remaining 76 words (9.48%) were non-equivalent due to factors such as missing entries in the lexicon, multiple possible root candidates, and cases where words beginning with mam-, man-, mang-, nam-, nan-, nang-, pam-, pan-, and pang- were actually base forms rather than derived ones. These findings highlight that while the rules are effective, additional linguistic and semantic considerations are necessary for full accuracy.

Table 13: Examples of Correct Assimilation and Reduction

Assimilated and Reduced Form	Root Word	Meaning of Assimilated and Reduced Form	Meaning of Root Word	Type of Assimilation and Reduction
mamasahi	pasahi	To pay the fare for riding a vehicle.	Transportation fare, the payment for a ride.	Rule 1: maNg- + /p,b/ → ma- + m + (word minus initial consonant)
manulisan	tulisan	To rob, to steal.	Robber or bandit, one who takes others possessions by force or intimidation.	Rule 2: maNg- + /t,d,s/ → ma- + n + (word minus initial consonant)
mangampanya	kampanya	To campaign, to go from house to house to convince of something.	A campaign for something.	Rule 3: maNg- + /k,g,'/ → maNg- + (word minus initial consonant or apostrophe)
mangaso	aso	To hunt using a dog.	A dog.	Rule 4: maNg- + /a, e, i, o, u/ → maNg- + (word)
mangyabi	yabi	To do something at night.	Night, the time when it is dark, without the sun.	Rule 4: maNg- + /y, w, l, r, m, n, ng/ → maNg- + (word)

Table 13 presents instances where the rules of assimilation and reduction were correctly applied to derive the root word. For example, *mamasahi* comes from *pasahi* ("fare for a ride"), where the prefix *maNg*- attaches to a root beginning with /p/, causing /ng/ to assimilate into /m/ and the initial consonant to drop. Similarly, *manulisan* ("to rob") derives from *tulisan* ("robber") through the same process, but with /t, d, s/ roots assimilating into /n/. In *mangampanya* from *kampanya* ("campaign"), the prefix preserves /ng/ when the root begins with /k/ or /g/. Other cases, such as *mangaso* from *aso* ("dog") and *mangyabi* from *yabi* ("night"), show how the prefix is retained without reduction when roots begin with vowels or approximants (y, w, l, r, m, n, ng).

These examples confirm that the linguistic rules governing assimilation and reduction are effective in explaining the formation of many derived words in Bolinao. The root and the derived forms maintain both morphological consistency and semantic alignment. In other words, the reconstructed root corresponds to the intended meaning of the assimilated form, demonstrating the reliability of the rules when applied to regular word formation processes.

Table 14: Examples of Incorrect Assimilation and Reduction

Assimilated and Reduced Form	Root Word	Meaning of Dissimulated Form	Meaning of Root Word	Type of Assimilation and Reduction
mama	ba	An intimate name for mother.	Equivalent to Booh! in play, To scare when used with babies.	Rule 1: maNg- + /p,b/ → ma- + m + (word minus initial consonant)
mani	ti	Peanut.	This specific.	Rule 2: maNg- + /t,d,s/ → ma- + n + (word minus initial consonant)
mangapo	gapo	To have grandchildren.	At all, an emphatic used with a negative limiting the degree.	Rule 3: maNg- + /k,g,'/ → maNg- + (word minus initial consonant or apostrophe)

mangaso	kaso	To hunt using a dog.	A court case, the matter that is brought before a legal body such as a judge.	Rule 4: maNg- + /a, e, i, o, u/ → maNg- + (word)
pansit'	tsit	Dish containing long fine rice noodles and vegetables, usually used for a snack or a special occasion..	Cheat.	Rule 4: maNg- + /t,d,s/ → ma- + n + (word minus initial consonant)

Table 14 shows cases where assimilation and reduction produced reconstructed roots that did not align semantically with their assimilated forms. For example, *mama* was derived from *ba* (“to scare in play”), but its actual meaning is “an intimate name for mother.” Similarly, *mani* (“peanut”) was tied to *ti*, which does not represent its true root. These errors reflect how the rules can sometimes undo forms incorrectly, especially when the assimilated and reduced word is itself a base form rather than a derived one. These are words beginning with mam-, man-, mang-, nam-, nan-, nang-, pam-, pan-, and pang- sometimes existed as base forms, but the rules mistakenly undid them.

Other examples reveal ambiguity in root selection. *Mangaso* could be reconstructed from *aso* (“dog”) or *kaso* (“case”), but only the former is semantically correct when meaning “to hunt with a dog.” Likewise, *pansit’* was derived from *tsit*, meaning “cheat,” but in reality, *pansit’* refers to a dish of noodles and vegetables. These cases underscore the limitations of a purely rule-based approach: while the mechanical process may generate plausible forms, it cannot always guarantee semantic accuracy. These findings confirm that while assimilation and reduction rules are effective for most cases, linguistic context and semantic checking are necessary to achieve full accuracy.

3.3.5 Pronoun Changes

Table 15: Summary of Rootword Extraction Results

Description	Count	Percentage
Total root words identified	28	100.0%
Semantically equivalent forms	28	100.0%
Non-equivalent forms	0	0.0%

Based on Table 15, all 28 pronouns in the dataset were successfully mapped to a corresponding root form using the rule-based stemmer. Every extracted root form preserved the semantic meaning of its original pronoun. This shows that the transformation process (e.g., prefix-stripping, normalization, or retention) did not result in loss or distortion of meaning. Therefore, no cases were found where the root candidate diverged semantically from the original pronoun. This further validates the effectiveness of the rules, suggesting that the algorithm performed consistently and without incorrect mappings.

Table 16: Examples of Pronoun Changes

Original Word	Root Word	Meaning of Original Word	Meaning of Root Word	Type of Pronoun Change
ansain	sain	Identifies a specific grouping of things close to the hearer when preceded by "no".	That group or activity near you, the hearer or an activity that happened previously.	Rule 1
ansaraytaw	saraytaw	What are those plural far.	Those specific persons, things or events far from the current context of the speaker.	Rule 1
inin	inin	Ah.... An expression used for something one cannot immediately recall or exactly express and used as a time lapse in speech.	Ah.... An expression used for something one cannot immediately recall or exactly express and used as a time lapse in speech.	Rule 3
taw	taw	That specific which is far in the context.	That specific which is far in the context.	Rule 3

Table 16 shows how the pronoun normalization process applies Rule 1 (Prefix-Stripping) and Rule 3 (Retainment) to selected Bolinao pronouns in order to obtain their canonical words while preserving grammatical meaning.

As we can see, the first two rows show interrogatives with the *ans-* prefix removed. Many interrogatives appear with the an- or ans- prefix. Therefore, the stemmer applies a prefix-stripping rule in order to undo these forms. This makes sure that all interrogatives are normalized to their root form without redundant morphophonemic material.

On the other hand, the third and fourth rows show the indefinites and expletives that generally retain their root forms in surface realizations. For indefinites, it is commonly expressed through reduplication, and since it is semantically integral rather than a surface alternation, these forms remain unchanged in terms of their root extraction. Undoing the reduplication would just obscure their indefinite meaning rather than normalize it, therefore, the stemmer preserves them as they are. For expletives, they similarly show no significant morphophonemic alternations. Therefore, their surface form is retained. However, the items here are still part of the core pronoun system of Bolinao alongside the other classes. Their inclusion is done in order to ensure that the pronoun inventory would be comprehensive and consistent, rather than excluding categories just because no change is detected in their root form.

Additionally, the absence of Rule 2 was observed as the dictionary-based normalization of *moy-/mod-* are not executed because no demonstratives with these prefixes appear in the lexicon. On the other hand, Rule 4 was done automatically at the beginning as pronouns under the Possessive Pronouns category in the lexicon were excluded as the said items differ significantly from the other pronoun classes that are included in the stemmer. Rather than working as core pronouns with root morphemes that are subject to prefixation or morphophonemic alternations, they act more like grammatical particles, linkers, or borrowed content words.

3.3.6 Number of Mass Indication

Table 17: Summary of Rootword Extraction Results

Description	Count	Percentage
Total root words identified	177	100.00%
Semantically equivalent forms	154	87.01%
Non-equivalent forms	23	12.99%

Based on Table 17, the detection process identified 177 root words, which were then evaluated for semantic equivalence with their original forms. Of these, 154 (87.01%) preserved meaning, showing that the processing effectively captured valid rootword correspondences. The remaining 23 (12.99%) did not retain equivalence, reflecting cases where extraction produced unattested or semantically unrelated forms.

Table 18: Examples of Correct Number of Mass Indication

Original Word	Root Word	Meaning of Original Word	Meaning of Root Word	Type of Mass Indication
babatag	batag	A banana grove.	Banana.	Rule 1: /baba/ → /ba/
bibingkol	bingkol	A rough area filled with uneven mounds of dirt.	Large clumps of dirt from plowing or digging.	Rule 1: /bibi/ → /bi/
dadayami	dayami	A place where there is straw.	The long rice stalk cut with anagodwith the rice and leaves.	Rule 1: /dada/ → /da/
kakawayan	kawayan	A bunch or an area of bamboo.	Bamboo.	Rule 1: /kaka/ → /ka/
tutupa'	tupa'	Muddy places.	Mud.	Rule 1: /tutu/ → /tu/

The examples in Table 18 illustrate cases where the detection of number or mass indication through reduplication produces root words that remain semantically equivalent to their original forms. Each word shows the systematic removal of an initial reduplicated CV sequence (Rule 1), yielding a valid root entry in the lexicon. For instance, babatag ("a banana grove") reduces to batag ("banana"), demonstrating how reduplication marks collectivity without obscuring the root meaning. These cases confirm that reduplication operates as a regular and predictable process in Bolinao, allowing surface forms to be reduced to their roots while preserving meaning.

Table 19: Examples of Incorrect Number of Mass Indication

Original Word	Root Word	Meaning of Original Word	Meaning of Root Word	Type of Mass Indication
kikiras	kiras	The edge of a dangerous cliff, gorge or canyon with numerous rocky areas.	A muffled crackling sound such as the rustling sound of leaves.	Rule 1: /kiki/ → /ki/
lalakwan	lakwan	A walkway.	The act of leaving someplace.	Rule 1: /lala/ → /la/
mama'	ma'	Combination of lime, betel nut, and tobacco that is chewed.	The moo of a cow or carabao.	Rule 1: /mama/ → /ma/
mamairot	mairot	To have something fitted tightly.	To be strict or adamant about something.	Rule 1: /mama/ → /ma/
tutukduan	tukduan	A base for something.	To step or stand or place on something solid such as a base, footings, foundation or reasoning.	Rule 1: /tutu/ → /tu/

The cases shown in Table 19 illustrate instances where reduplication produces rootword candidates that diverge semantically from their original forms or fail to reflect true mass indication. Although the detection rules correctly reduced the initial CV sequences (Rule 1), the resulting roots did not always preserve the intended meaning. For example, kikiras ("the edge of a dangerous cliff") reduces to kiras ("a muffled crackling sound"), yielding a root that functions as an unrelated lexical item. Similarly, tutukduan ("a base for something") reduces to tukduan ("to step or stand on a base or foundation"), which is semantically related but does not express plurality or collectivity since the original form is already singular. These cases demonstrate that while reduplication detection is systematic, it can generate false positives either through semantic mismatch or through forms that are not true indicators of number or mass. This highlights the need for lexicon filtering and semantic evaluation in verifying rootword candidates.

The analysis demonstrates that reduplication effectively signals number or mass and, in many cases, allows accurate recovery of the rootword while preserving its meaning. Nonetheless, certain forms revealed semantic divergence, indicating that reduplication alone is not always reliable without further contextual or lexical validation. Moreover, other morphological markers such as gemination and infixation, which are also recognized indicators of plurality or collectivity, were absent from the dataset and therefore could not be assessed. This highlights both the utility and the limitations of the current approach in evaluating number or mass indication in Bolinao.

3.3.7 Gemination and Reduction

Table 20: Transformation Counts

Description	Count	Percentage
Total root words identified	122	100.0%
Gemination	47	38.5%
Reduction	63	51.6%
Both Applied	12	9.8%

Table 20 showed that vowel reduction was the most frequently applied transformation, accounting for 63 of 122 reconstructed entries, or 51.6 percent of the dataset. This reflects the prominence of consonant clustering in Bolinao, particularly in environments affected by vowel syncopation, infixation, or reduplication. Restoration of vowels based on consonantal context yielded many valid root matches, confirming the utility of phonologically guided reconstruction.

Gemination collapse was applied to 47 entries, representing 38.5 percent. These transformations typically involved expressive or emphatic forms in which doubled consonants were simplified. The results suggest that gemination is a productive phonological feature in Bolinao, and its reversal can reliably recover root forms when applied to appropriate contexts.

Twelve entries (9.8 percent) required gemination and reduction, indicating layered phonological processes. These cases demonstrate the interaction between consonant doubling and vowel loss, often occurring within morphologically complex or inflected forms.

The table showed that rule-based reconstruction, when applied to a lexicon-aligned dataset, can effectively recover root forms and preserve semantic equivalence in a substantial portion of entries. The distribution of transformation types supports the validity of the reconstruction logic and highlights the phonological regularities present in Bolinao word formation.

Table 21: Proper application of Reduction

Reduced Form	Root Word	Meaning of Reduced Form	Meaning of Root Word	Type of Reduction
gusgusen	gusagusen	To scrub something especially hard to remove	To scrub something aggressively	Vowel restoration (sg → sag)
gusgos	gusagos	A material used for cleaning something like the skin or body.	The act of scraping something, usually to remove something.	Vowel restoration (sg → sag)
kaallakyan	kaalakyan	The size of something	The comparative size or measurement of something, its magnitude	Vowel restoration (lk → lak)
dayday	dayaday	To spread something out.	To be laid out or hung out to dry.	Vowel restoration (bl → bil)
kumbit	kumibit	To become rigid and unyielding	To glare angrily at someone	Vowel restoration (mb → mib)

Table 21 shows that reduction was applied successfully across multiple entries, restoring vowels within consonant clusters based on predictable phonological environments. Each reconstructed root matched a valid entry in the Bolinao lexicon and retained semantic equivalence with the original surface form. These results confirm that vowel restoration, when guided by consonantal context, is a reliable method for recovering root words affected by syncopation or infixation. The consistency of meaning across both forms supports the linguistic validity of the reduction process in Bolinao.

Table 22: Reduction with Semantic Divergence

Reduced Form	Root Word	Meaning of Reduced Form	Meaning of Root Word	Type of Reduction
igusgos	igusagos	To rub gently as in washing parts of the body or soaping a washcloth	Do something in anger	Vowel restoration (sg → sag)
pikwen	pikawen	To weigh something	To pierce something with a spear-like object	Vowel restoration (kw → kaw)
kulkol	kulakol	A dispute that is brought to court	To be drained of any energy, even more so than napagal, exhausted	Vowel restoration (lk → lak)

blangen	bilangen	To split wood with the grain using an ax or heavy bolo	To calculate or count out something	Vowel restoration (bl → bil)
lumtak	lumitak	To come to have a split because of lack of water or such as a seed germinating	To click the tongue with the mouth wide open	Vowel restoration (mt → mit)

Table 22 shows that although reduction was applied to these words, the meaning of the reduced form does not match the meaning of the root word. This means that the transformation changed how the word is used or understood. In some cases, the reduced word became more specific, more general, or took on a different role in the sentence. These examples show that reduction does not always give back the original meaning, and extra care is needed when using it to find root words.

Table 23: Reduction Semantic Validation

Description	Count	Percentage
Total words with reduction applied	63	100.0%
Properly applied	28	44.4%
Improperly applied	35	55.6%

Table 23 shows that reduction was applied to 63 entries in total. Of these, 28 entries (44.4 percent) retained semantic equivalence between the reduced form and the confirmed root, indicating successful vowel restoration. These cases reflect predictable phonological environments where reduction did not distort meaning.

However, 35 entries (55.6 percent) showed semantic divergence, suggesting that the reduction process either introduced lexical ambiguity or resulted in a derived form with a distinct meaning. This highlights the need for semantic validation when applying reduction rules

Table 24: Proper application of Gemination

Geminated Form	Root Word	Meaning of Geminated Form	Meaning of Root Word	Type of Gemination
mammuno'	mamuno'	A prayer leader	To lead responsively	Consonant collapse (mm → m)
magganansya	maganansya	To gain a profit	The profit from conducting business	Consonant collapse (gg → g)
maggwardya	magwardya	To guard or protect something as a responsibility or job	To be guarding something	Consonant collapse (gg → g)
mammapas	mamapas	The substance that brings about fading	To fade	Consonant collapse (mm → m)
mammarang	mamarang	An evil spirit that leads a person astray	To lead astray	Consonant collapse (mm → m)

Table 24 shows that gemination was applied correctly in these examples, where doubled consonants were simplified to recover root forms. Each transformation preserved the original meaning and matched a valid entry in the Bolinao lexicon. This confirms that gemination collapse is a reliable method for reconstructing root words, especially in expressive or emphatic forms. The consistent semantic match across these entries supports its use in rule-based modeling for Bolinao.

Table 25: Gemination with Semantic Divergence

Geminated Form	Root Word	Meaning of Geminated Form	Meaning of Root Word	Type of Gemination
agwa	agawa	Perfume water	Diligence combined with promptness	Consonant collapse (gw → g)
alla	ala	Watch out! (warning)	Go! (command)	Consonant collapse (ll → l)
arkan	arakan	To kiss or touch with the cheek or lips	To flock to something	Consonant collapse (rk → r)
bakbak	bakabak	A fight between dogs or pigs	A dry wood termite	Consonant collapse (kk → k)
barbar	barabar	To soak something in liquid	Stick for barbecuing	Consonant collapse (rb → r)

Table 25 shows that gemination was applied to these words, but the meaning of the geminated form does not match the meaning of the root word. This means the transformation changed the word's meaning or created a new word entirely. In some cases, the geminated word became more expressive, idiomatic, or took on a different role. These examples show that gemination does not always help recover the original meaning, and it should be used carefully when identifying root words.

Table 26: Gemination Semantic Validation

Description	Count	Percentage
Total words with gemination applied	47	100.0%
Properly applied (same meaning)	22	46.8%

Improperly applied (different meaning)	25	53.2%

Table 26 shows that gemination was applied to 47 words. Out of these, 22 words kept the same meaning as their root, which means the transformation worked well. However, 25 words changed meaning after gemination. This means that gemination sometimes creates new words or shifts the meaning, so it should be used carefully when trying to find the original root.

The analysis of both gemination and reduction processes in the Bolinao dataset reveals a consistent pattern in semantic outcomes. While a number of reconstructed forms successfully preserved the meaning of their root candidates, a substantial portion did not. Specifically, reduction was applied to sixty-three entries, of which only twenty-eight (44.4%) retained semantic equivalence with their confirmed roots. Similarly, gemination was applied to forty-seven entries, with only twenty-two (46.8%) preserving the original meaning.

These figures indicate that more than half of the transformed words in each category resulted in semantic divergence from their root forms. In such cases, the transformed word cannot be considered usable for root recovery in a strict linguistic sense. Rather than functioning as phonological variants, these forms represent derived lexical items, words that have undergone semantic, functional, or grammatical evolution beyond their original structure. This phenomenon is referred to as word derivation, and it plays a significant role in the development of the language.

4. CONCLUSION AND FUTURE WORK

This research directly addresses the critical lack of digital resources for the Bolinao language which is a key barrier to its inclusion in the modern NLP landscape. Through the systematic scraping, cleaning, and standardization of data from the Webonary platform, we have produced the largest public accessible, NLP-ready for Bolinao, comprising a total of 20,308 unique lemmas. The vital step of mapping the custom part-of-speech tags to the Universal Part-of-Speech (UPOS) standard ensures that this lexicon is immediately a usable resource for the global NLP community. Furthermore, our preliminary computational analysis of Bolinao morphophonemic processes provides the necessary groundwork for developing more advanced Bolinao language processing tools. The lexicon's high coverage rate that was validated against the secondary corpora of songs and stories, confirms its value as a robust foundation for future research and development of applications.

The lexicon opens several promising avenues for future work. The findings and resources from this journal paper report can be extended in the following ways:

1. The rule-based analysis conducted in this paper can be operationalized to build a fully functional stemmer for Bolinao.
2. The lexicon can be directly integrated into applications that serve the Bolinao speakers. An e-governance tool for official documents written in Bolinao is now a feasible objective.
3. While the lexicon is comprehensive, it can be further enhanced by incorporating data from more diverse sources such as social media data and digitized literature. This would scale its coverage of neologisms, slang, and dialectal variations.

5. DISCLAIMER

Throughout the development of this journal paper report, the team utilized large language models as a supplementary tool to streamline tasks such as technical description and refinement of text. The final research design, analysis, and conclusions remain with the responsibility of the authors. All content generated by the large language models undergo careful review, validation, and revision to guarantee it met the academic standards that reflect the team's original work.

6. REFERENCES

[1] Stuart Webb. 2021. Research Investigating Lexical Coverage and Lexical Profiling: What We Know, What We Don’t Know, and What Needs to be Examined. 33, 2 (2021). Retrieved from https://files.eric.ed.gov/fulltext/EJ1316858.pdf

[2] Jhonabell Alejan, Jane Irish, E Ayop, Jescilla Alojado, Pearl Abatayo, Krishna Sol, Rene Abacahin, and Bonifacio. 2021. Heritage Language Maintenance and Revitalization: Evaluating the Language Endangerment among the Indigenous Languages in Bukidnon, Philippines. Retrieved from https://files.eric.ed.gov/fulltext/ED617996.pdf

[3] Khang Nhut Lam, Feras Al Tarouti, and Jugal Kalita. 2014. Creating Lexical Resources for Endangered Languages. arXiv (Cornell University) (January 2014). DOI:https://doi.org/10.3115/v1/w14-2207

[4] Gary C. Persons. 1978. BOLINAO: A PRELIMINARY PHONEMIC STATEMENT. Retrieved September 13, 2025 from https://www.sil.org/system/files/reapdata/86/96/90/869690144061260043347839124796287677 1/smk_A_Preliminary_Phonemic_Statement_1978.PDF

[5] Gary C. Persons. 1980. Bolinao morphophonemics. SIL Global. Retrieved September 13, 2025 from https://www.sil.org/resources/archives/90565

[6] Miriam Corris, Christopher Manning, Susan Poetsch, and Jane Simpson. 1999. Perth Congress of the Applied Linguistics Association of Australia. Retrieved September 26, 2025 from https://nlp.stanford.edu/pubs/corris1999dictionaries.pdf

[7] Ulrike Mosel. 2004. Dictionary making in endangered speech communities. Language Documentation and Description 2, 0 (July 2004). DOI:https://doi.org/10.25894/ldd289

[8] Ulugbek Nurmukhamedov and Stuart Webb. 2019. Lexical coverage and profiling. Language Teaching 52, 02 (April 2019), 188–200. DOI:https://doi.org/10.1017/s0261444819000028

[9] Dan Jurafsky and James H. Martin. 2025. Speech and Language Processing. Stanford.edu. Retrieved from https://web.stanford.edu/~jurafsky/slp3/

[10] Kenneth R Beesley and Lauri Karttunen. 2003. Finite-State Morphology. Retrieved from https://www.researchgate.net/publication/37705086_Finite-State_Morphology

[11] Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing 4, 1 (January 2007), 1–34. DOI:https://doi.org/10.1145/1187415.1187418

[12] Ryan Cotterell, Ekaterina Vylomova, Huda Khayrallah, Christo Kirov, and David Yarowsky. 2017. Paradigm Completion for Derivational Morphology. DOI:https://doi.org/10.48550/arXiv.1708.09151