Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Mapping Morphological Patterns: A Framework for Rinconada Bikol Language Morphological Analysis and Stemming

Tiffany Lyn PANDES[ab] & Joshua MARTINEZ[a]
[a]*College of Computer Studies, Ateneo De Naga University, Philippines*
[b]*College of Computer Studies, Camarines Sur Polytechnic Colleges, Philippines*
tlpandes@gbox.adnu.edu.ph
joshuamartinez@gbox.adnu.edu.ph

Abstract: Natural Language Processing (NLP), a subfield of Artificial Intelligence (AI), has gained traction in management research, particularly linguistics. However, only High-resource language is being established in NLP. This paper aims to analyze morphological patterns of Low-resource languages with limited linguistic data and resources available for NLP tasks such as Rinconada Bikol Language (RBL). This paper proposed a framework suited for RBL as the approach to developing the RBL Morphological Analyzer. This paper utilized the framework and evaluated it using Morphological Accuracy, revealing an impressive 90% accuracy in identifying correct analysis and stemming. The system's precision stands at 0.90, with a perfect recall of 1.00, resulting in an F1 score of 0.95. This high level of performance indicates the system's strong ability to recognize morphological features and patterns within the dataset effectively. The findings also reveal that the framework could also accurately analyze the morphological structure of RBL sentences.

Keywords: NLP, morphology, morphological analyzer, stemmer, RBL

1. Introduction

Natural language processing (NLP) has garnered significant attention in the field of management research, particularly within the realm of linguistics, owing to its automated capacity for analyzing and comprehending human language(Devi & Purkayastha, 2018; Meurers, 2021). While many NLP systems for English overlook word morphology, it is crucial in numerous other languages, emphasizing the importance of understanding human knowledge of morphology in cognitive science(Aronoff et al., 2005; Bhanvadia et al., 2022). Effectively addressing morphology can help mitigate sparse data challenges in NLP, especially for low-resource languages(Parhat et al., 2019; Schäfer et al., 2022). The complexity of morphology, including inflection, derivation, and compounding, allows languages to convey extensive information in a single word, but it also presents challenges for NLP systems(Koehn et al., 2003; Liu et al., 2022). Moreover, the study of Rinconada Bikol Language (RBL), spoken by approximately 500,000 individuals, is crucial due to its secluded location and small population, highlighting the importance of developing NLP tools for such minority languages(Arrivillaga & Feliciangeli, 2001; Prenner et al., 2022).

Morphology handling is essential not only for NLP tasks but also for preserving endangered languages. While serving as NLP tools, rule-based methods also act as machine-readable documentation of languages, ensuring accuracy, especially in endangered languages (Gamallo et al., 2019; Schmidt-Schauß & Sabel, 2020). However, the continuous evolution of languages necessitates the integration of neural models, which can learn to generalize principles for words not in their vocabulary, addressing the challenge of keeping up with the rate of linguistic changes (Ruder et al., 2022; Vásquez et al., 2022). Therefore, the development of an artificially intelligent method for identifying morphological patterns, as proposed in the paper "Mapping Morphological Patterns: A Framework for Rinconada Bikol

Language Morphological Analysis and Stemming," holds significant promise for Language Corpus Construction, machine translation, and stemming applications for various languages, including Filipino and minority languages such as RBL (De Torre & Gonong, 2020; Swathi & Jayashree, 2020). This addresses the following research questions: (1) How can the design framework and prototype for morphological analysis be applied to the RBL dataset? (2) What is the effectiveness of the RBL Morphological Analysis framework and prototype?

2. Related Work

2.1 Intro to NLP, Morphological Analysis and Stemming

Morphological analysis and stemming are important techniques in natural language processing, allowing for the decomposition of words into their constituent parts and reducing words to their base or root form, respectively. In recent years, morphological analysis and stemming advancements(Childs, 2014) have focused on improving these processes' accuracy and efficiency and expanding their capabilities to handle complex linguistic phenomena.

Morphological analyzers identify root words and their features by extracting affixes, which can be prefixes, infixes, suffixes, or circumfixes. Each type of affix attaches to a word differently, adding to its meaning. These analyzers are trained on data that has been manually tagged to indicate the boundaries between segments and their parts of speech. This process is especially important for languages like Japanese and Chinese, which lack clear word boundaries, requiring detailed word segmentation for analysis (Tolmachev et al., 2019). Morphological stemming is crucial to NLP preprocessing. Applying machine learning to real-world datasets requires a lot of preprocessing, from converting formats to tokenizing and stemming text (Tong et al., 2017).

2.2 Rinconada Bikol Language (RBL)

Rinconada Bikol is a minority language spoken in the Philippines, belonging to the Austronesian language family (Helwig et al., n.d.). It is endangered due to urbanization, migration, and the growing use of Filipino and English in educational and governmental contexts. Efforts have been made to document and preserve the language, but more action is needed to ensure its long-term survival. It uses affixes to convey diverse grammatical functions and exhibits a complex verb conjugation system (Lobel, 2004), with multiple forms for different tenses, aspects, and moods.

3. Methodology

3.1 Conceptual Framework

Figure 1 depicts the study's conceptual framework, which includes five main phases: dataset collection, dataset preparation, framework design and development, framework evaluation, and RBL morphological analysis. The phases are interconnected with feedback loops, ensuring continuous improvement through iterative processes.

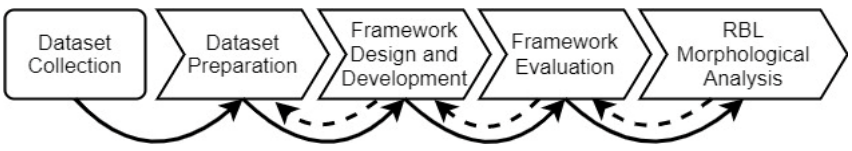

Figure 1. Conceptual Framework

3.1.1 Dataset Collection

This is the initial phase, where raw data relevant to the study or project is gathered. The data could come from various sources such as manual to crowdsourcing, engaging native speakers and language enthusiasts to contribute data. The types of data collected include textual data and annotated data. The volume of data collected should be substantial, with a minimum target of 3,000 words for low-resource languages. Quality assurance processes ensure data accuracy, and the collected datasets are relevant to the specific context of resource-based learning. This phase feeds directly into the Dataset Preparation phase.

3.1.2 Dataset Preparation
In this phase, the collected data is cleaned, transformed, and organized to ensure it is in a usable format. This may involve handling missing values, normalization, and splitting the dataset into training and testing sets. Prepared data is then passed on to the Framework Design and Development phase. If issues arise during this phase or later stages, the data may need to be re-prepared, which would loop back to this phase.

3.1.3 Framework Design and Development
Here, the actual framework or model is designed and developed. This involves selecting specific algorithms, such as the Porter Stemming Algorithm, and defining architectures that may include rule-based systems. The development process includes coding and integrating different components using programming tools like Python and development environments like Jupyter Notebook or Native Python. The design of the framework is heavily dependent on the preparation of the dataset. Once designed, it is passed on to the Framework Evaluation phase for testing and validation. If the framework does not perform as expected, adjustments may need to be made, requiring a loop back to this phase.

3.1.4 Framework Evaluation
During this phase, the framework is tested and validated using specific evaluation metrics like accuracy, recall, precision, and F1 score. Test datasets are carefully chosen to cover various linguistic scenarios. The evaluation process includes steps like cross-validation and confusion matrix analysis using tools like scikit-learn. The framework undergoes thorough evaluation to ensure its reliability and functionality, and any issues are identified and addressed. Depending on the results, the framework may need refinement or redevelopment, or changes may be required in dataset preparation, leading back to the Framework Design and Development phase.

3.1.5 RBL Morphological Analysis
The framework implementation occurs during the RBL morphological analysis phase, involving iterative processes and a comprehensive exploration of morphological features. Specific steps include data preprocessing, feature extraction, and analysis, employing techniques like stemming and lemmatization to extract morphological features. Insights gained from this analysis may lead to further refinement of the dataset, framework design, or evaluation, resulting in iterative loops back to the earlier stages.

4. Results and Discussion

4.1 Morphological Framework

The framework is based on linguistic rules specific to the morphological patterns in RBL. Special attention is given to ensure accuracy and adaptability to the language's unique morphological patterns. These rules serve as guiding principles for morphological analysis.

4.1.1 Pre-fix rule
The prefix rule in grammar involves adding affixes to the beginning of a root word to modify its meaning or create new words.

Table 1. Prefix Rule (Unlapi)

Prefix	+Root Word	Word
sing	+linig	=sinlinig
pan	+linig	=panlinig
taga	+manda	=tagamanda

Table 1 shows the Prefix rule; for example, the prefix *"sing-"* can be added to the word *"linig"* (clean) to create *"sinlinig"* which means "just clean." Prefixes can also indicate purpose, such as the prefix *"pang-,"* which can be added to *"linig"* to create *"panlinig,"* meaning "for cleaning."

4.1.2 Infix rule
The infix rule involves adding affixes to the middle of a root word, allowing for creating new words and modifying existing ones by inserting prefixes or suffixes into the word stem.

Infixation is a common feature of many languages and can change the meaning of a word, alter its grammatical function, or create entirely new words.

Table 2. Infix Rule (Gitlapi)

Infix	+Root Word	Word
-in-	+singkil	=siningkil
-um-	+dungkal	=dumungkal

Table 2 shows the infix rule (gitlapi), a rule in the Filipino language that allows you to add words to strings. The infix rule is applied by placing the infix between the syllables of the root word. In the first two examples, the infix "*-in-"* is used to add the root word *"singkil"* (stumble). The resulting word is *"siningkil"* (to strike the foot against something). In the fourth example, the infix *-um-* is used for the root word *"dungkal"* (stub). The resulting word is" dumungkal*"* (to stub).

4.1.3 Suffix rule

The suffix rule involves adding an affix to the end of a root word to modify its meaning, create new words, or form different parts of speech.

Table 3. Suffix Rule (Hulapi)

Suffix	+Root Word	Word
-on	+sariwa	=sariwaon
-an	+iyan	=iyanan
-hon	+bagu	=baguhon

Table 3 shows the suffix rule in the RBL. The suffix *"-on"* is added to root words that describe a quality or state to form nouns denoting things or people with that quality or state. For example, the root word *"sariwa"* means "fresh". When the suffix"-*on"* is added to it, the resulting word *"sariwaon"* means "something fresh," such as a fresh fish or a fresh fruit. The suffix *"-an"* is added to root words denoting actions or events to form nouns denoting places where those actions or events occur. For example, the root word*" iyan"* means" go." When the suffix *"-an"* is added to it, the resulting word *"iyanan"* means "to go." The suffix"-*hon"* is added to root words that denote verbs to form adjectives that describe people or things that are good or suitable for doing those verbs.

4.1.4 Prefix, Infix, Suffix rule

Prefixes are added to the beginning of a word, infixes are inserted in the middle, and suffixes are added to the end. This combination of components creates a complex word structure that forms the basis of many languages worldwide.

Table 7. Prefix, Infix, Suffix Rule

Prefix, infix, suffix	+Root Word	Word
pag- -um- -on	+sikap	=pagsumikapon
ipag- -um- -an	+siyak	=ipagsumiyakan
mag- -in- -an	+pusta	=magpinustan

Table 7 shows the prefix, infix, and suffix rules in Filipino. In the first example, the prefix *"pag-"*,the infix *"-um-"*, and the suffix *"-an"* are added to the root word *"sikap"* to form the word "*pagsumikapon"*, which means "to be hardworking" or "to make an effort." In the second example, the prefix *ipag-*, the infix *-um-*, and the suffix *-an* are added to the root word *"siyak"* to form the word *"ipagsumiyakan"*, which means" to make an example of someone."

4.2 Prototype

The RBL prototype processes input text in its native language using tokenization and segmentation. This breaks down the text into morphemes, which are further analyzed using a morphological decomposition module to identify each word's features and patterns.

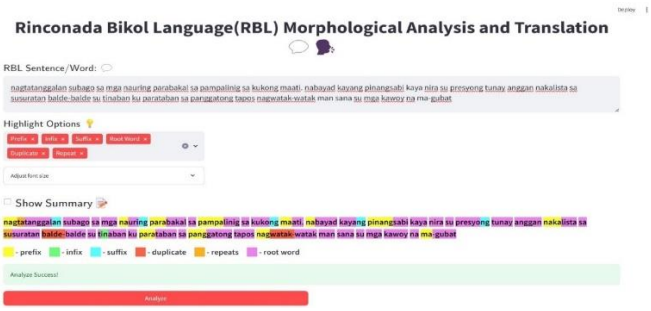

Figure 2. Morphological Analysis Prototype

The Prototype UI in Figure 3 is a tool for analyzing sentences in the RBL language, spoken in the Philippines. It breaks down words into morphemes, which are constituent parts of a word, such as roots or affixes. The tool identifies stop words, prefixes, infixes, and suffixes, and then determines the root word for each word.

4.3 Evaluation

Table 8. Morphology Accuracy

Total Validation Size	**500**
Number of Correct Morphology	**450**
Accuracy	**90%**
Precision	**0.90**
Recall	**1.00**
F1 Score	**0.95**

In Table 8, the system achieved a 90.00% accuracy rate, accurately identifying 450 out of 500 samples in morphological analysis. It demonstrated a precision of 0.90 and a perfect recall of 1.00, resulting in an F1 score of 0.95. However, with a 10% error rate, there are evident areas for enhancement. Delving into these errors, particularly in identifying complex morphological patterns, could bolster the system's capabilities. Accurate morphological analysis has widespread implications, including machine translation, where it can enhance precision by preserving word structures and meanings. A substantial sample size (500) for the evaluation underscores a comprehensive testing and validation process, instilling confidence in the system's performance. In summary, while the system demonstrates strong performance in morphological analysis, there is still potential for improvement to elevate its accuracy and utility across various tasks.

5. Conclusion

This paper further concludes the field by evaluating the accuracy of morphological analysis, revealing an impressive 90% accuracy in identifying correct stemming from a validation size of 500 samples. The system's precision stands at 0.90, with a perfect recall of 1.00, resulting in an F1 score of 0.95. This highlights the framework's potential to effectively recognize morphological features, which is crucial for stemming tasks. Despite the strong performance in morphological analysis and translation tasks, the study acknowledges a 10% error rate in morphological analysis, suggesting room for further improvement. It proposes that analyzing errors and identifying challenging morphological patterns could enhance the system's capabilities. Given the complexity of RBL morphology, developing a robust Morphological Analyzer is crucial. This tool should be capable of accurately analyzing the morphological structure of RBL sentences, thereby aiding the translation process. Continuous development and refinement of the Morphological Analyzer, based on feedback from its application in the pipeline, will significantly enhance accuracy.

Acknowledgements

We thank the following: Ateneo De Naga University (ADNU) and our loved ones for their unselfish support in writing this paper.

References

Aronoff, M., Meir, I., & Sandler, W. (2005). The Paradox of Sign Language Morphology. *Language*. https://doi.org/10.1353/lan.2005.0043

Arrivillaga, J., & Feliciangeli, M. D. (2001). <i>Lutzomyia Pseudolongipalpis:</I>The First New Species Within The<i>longipalpis</I>(Diptera: Psychodidae: Phlebotominae) Complex From La Rinconada, Curarigua, Lara State, Venezuela. *Journal of Medical Entomology*. https://doi.org/10.1603/0022-2585-38.6.783

Bhanvadia, S. B., Saseendrakumar, B. R., Guo, J., Daniel, M., Lander, L., & Baxter, S. L. (2022). Evaluation of Bias in Medical Student Clinical Clerkship Evaluations Using Natural Language Processing. *Academic Medicine*. https://doi.org/10.1097/acm.0000000000004807

Childs, P. R. N. (2014). 3.7 Morphological Analysis. In *Mechanical Design Engineering Handbook*. Elsevier. https://app.knovel.com/hotlink/khtml/id:kt010SQBA1/mechanical-design-engineering/morphological-analysis

De Torre, R. G., & Gonong, G. O. (2020). A Phono-Lexicostatistical Analysis of Bikol-Sorsogon Varieties. *The Normal Lights*. https://doi.org/10.56278/tnl.v14i2.1655

Devi, M. I., & Purkayastha, B. S. (2018). Advancements on NLP Applications for Manipuri Language. *International Journal on Natural Language Computing*. https://doi.org/10.5121/ijnlc.2018.7505

Gamallo, P., Sotelo, S., Pichel Campos, J. R., & Artetxe, M. (2019). Contextualized Translations of Phrasal Verbs With Distributional Compositional Semantics and Monolingual Corpora. *Computational Linguistics*. https://doi.org/10.1162/coli_a_00353

Helwig, N. E., Hong, S., & Hsiao-wecksler, E. T. (n.d.). *Rinconada Bikol language*. https://en.wikipedia.org/wiki/Rinconada_Bikol_language

Koehn, P., Och, F. J., & Marcu, D. (2003). *Statistical Phrase-Based Translation*. https://doi.org/10.21236/ada461156

Liu, C., Ge, S., & Liu, H. (2022). *Toward Understanding Bias Correlations for Mitigation in NLP*. https://doi.org/10.48550/arxiv.2205.12391

Lobel, J. W. (2004). Old Bikol -um- vs. mag- and the loss of a morphological paradigm. *Oceanic Linguistics*, *43*(2), 469–497. https://doi.org/10.1353/ol.2005.0007

Meurers, D. (2021). *Natural Language Processing and Language Learning*. https://doi.org/10.1002/9781405198431.wbeal0858.pub2

Parhat, S., Ablimit, M., & Hamdulla, A. (2019). A Robust Morpheme Sequence and Convolutional Neural Network-Based Uyghur and Kazakh Short Text Classification. *Information*. https://doi.org/10.3390/info10120387

Prenner, J. A., Babii, H., & Robbes, R. (2022). *Can OpenAI's Codex Fix Bugs?* https://doi.org/10.1145/3524459.3527351

Ruder, S., Vulić, I., & Søgaard, A. (2022). *Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold*. https://doi.org/10.18653/v1/2022.findings-acl.184

Schäfer, H., Idrissi-Yaghir, A., Horn, P. A., & Friedrich, C. M. (2022). *Cross-Language Transfer of High-Quality Annotations: Combining Neural Machine Translation With Cross-Linguistic Span Alignment to Apply NER to Clinical Texts in a Low-Resource Language*. https://doi.org/10.18653/v1/2022.clinicalnlp-1.6

Schmidt-Schauß, M., & Sabel, D. (2020). Correctly Implementing Synchronous Message Passing in the Pi-Calculus by Concurrent Haskell's MVars. *Electronic Proceedings in Theoretical Computer Science*. https://doi.org/10.4204/eptcs.322.8

Swathi, S., & Jayashree, L. S. (2020). *Machine Translation Using Deep Learning: A Comparison*. https://doi.org/10.1007/978-3-030-24051-6_38

Tolmachev, A., Kawahara, D., & Kurohashi, S. (2019). Shrinking Japanese Morphological Analyzers With Neural Networks and Semi-supervised Learning. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2744–2755. https://doi.org/10.18653/v1/N19-1281

Tong, K., Soergel, D., Katsiapis, G., & Engineers, S. (2017). Preprocessing for Machine Learning with tf.transform. In *Google AI Blog*. Google Research. https://ai.googleblog.com/2017/02/preprocessing-for-machine-learning-with.html

Vásquez, J. G., Bel-Enguix, G., Andersen, S. W., & Ojeda Trueba, S. L. (2022). *HeteroCorpus: A Corpus for Heteronormative Language Detection*. https://doi.org/10.18653/v1/2022.gebnlp-1.23