PANG-KAT: A Dedicated Tokenizer for the Tagalog Language

Justin Louis L. Saavedra and Jaderick P. Pabico

JUNE 2025

2025 ICS UNDERGRADUATE RESEARCH SYMPOSIUM

INTRODUCTION

INTRODUCTION

Tokenization is a critical preprocessing step in many natural language processing (NLP) tasks, yet most general tokenization methods do not effectively recognize the unique grammatical features [2], named entities (NEs), and multi-word expressions (MWEs) [3] of a language.

INTRODUCTION

Low-resource languages are more prone to these issues [2], including the Tagalog language, which is the most spoken language in the Philippines, but lacks a suitable tokenizer for accurately processing its language resources to promote its further NLP advancements. These issues on tokenization is further complicated by code-switching [1].

OBJECTIVES

Develop a dedicated tokenizer for the Tagalog language to address its lack of a suitable tokenizer, which would also incorporate the recognition of Tagalog NEs and MWEs, and Tagalog-English (Taglish) tokenization in an attempt to make a general tokenizer for real-world Tagalog language processing applications.

OBJECTIVES

- To create a dictionary containing Tagalog NEs and multi-word expressions to facilitate their recognition in the tokenization process.
- To annotate a Taglish corpus into pre-segmented tokens, to be used for assessing the adaptability of the Tagalog tokenizer on Taglish data.
- To construct a hybrid rule and dictionary-based tokenizer for the Tagalog language
- To evaluate the tokenizer based on the following performance evaluation metrics:
  a. Accuracy
  b. Precision
  c. Recall
  d. F1-Score

METHODOLOGY

METHODOLOGY

To achieve its objectives, the study introduces PANG-KAT, a Python module developed as a hybrid rule and dictionary-based tokenizer due to the lack of pre-annotated Tagalog resources for a machine-learning approach [1].

DICTIONARY CREATION

To facilitate the recognition of Tagalog NEs and MWEs, a dictionary for PANG-KAT was manually created by extracting its vocabulary from various publicly available datasets and online resources [15-16] [25-32].

TWEETTAGLISH ANNOTATION

The author, in collaboration with two native Tagalog speaker annotators, annotated the TweetTaglish corpus [1] using Inside–Outside–Beginning (IOB) tagging, which was used in observing patterns in Tagalog and English (Taglish) code-switching that will be incorporated in the rule extraction phase to make PANG-KAT capable of Taglish tokenization.

RULE EXTRACTION

For PANG-KAT rules, these were manually extracted from two available Universal Dependencies (UD) treebanks in the Tagalog Language: Tagalog Reference Grammar (TRG) [18] and Ugnayan [8], cross-referenced with KWF's Manwal sa Masinop na Pagsulat [33], and on the TweetTaglish corpus [1].

Example of short and longer unit tokenization with their corresponding labels. PANGKAT's dictionary categories include MWEs and NEs such as organization (NE-ORG), location (NE-LOC), and person (NE-PER) entities. Non-NEs and MWEs are labelled as "W", which corresponds to a word.

PERFORMANCE EVALUATION

PANG-KAT's performance was evaluated through unit testing and external validation on the manually annotated NewsPH-NLI dataset [20] and May 2025 articles from Pilipino Star Ngayon [21] using the performance evaluation metrics of accuracy, precision, recall, and F1 Score.

RESULTS & DISCUSSION

DICTIONARY CREATION

CATEGORY	TOKEN COUNT
Location entities (NE-LOC)	48,383
Organization entities (NE-ORG)	3,564
Person Entities (NE-PER)	7,321
Multi-word Expressions (MWE)	3,294
PANG-KAT's Total Dictionary Size	**62,562**

The dictionary size and composition of PANG-KAT

TWEETTAGLISH ANNOTATION

LEGEND	COUNT
Sentences	1858
Labeled Tokens	34,424

The composition of the annotated TweetTaglish Corpus

RULE EXTRACTION

RULE	EXAMPLE
Intensification of word meaning	malungkot na malungkot, magandang maganda, darating at darating
Continuity of action	tumakbo nang tumakbo, palakas nang palakas
Repeating words	gaya-gaya, kani kanila, tumalon-talon, nagtuloy tuloy, ang lungkot-lungkot
Contractions	ako'y, dalawampu't apat, s'ya, '97, 'wag
Tagalog date and time	ika-15 ng Abril, a-kinse ng Abril, ika-tatlo ng hapon, alas tres ng hapon

The manually extracted rules of the Tagalog language

RULE EXTRACTION

RULE	EXAMPLE
Borrowed Spanish conjunction “y”	alas tres y medya, trenta y dos, Jose Protasio Rizal y Alonso Realonda
Daglat	Kgg. Pangilinan, G. Pangilinan, Pangilinan, PhD, DepEd
Spelled-out large Tagalog numbers	limampu’t isang libo, walong daang libo, sampung milyon, 11 bilyon
Beginning Markers for Taglish code-switching (Tagalog prefix + English word)	nagma-marites, ipa-consult, pagkabitter, napaka close

The manually extracted rules of the Tagalog language

RULE SEQUENCING

PANG-KAT operates on a linear, first-match-wins architecture, wherein when a token is labeled, it won’t be reconsidered by later rules. Thus, rule sequencing is crucial for ensuring accurate token labeling, which starts with more structured, Tagalog-specific rules, followed by general rules on the proper usage of punctuation, and the broadest dictionary lookups.

RULE SEQUENCING

Token labels are only updated based on specific triggers that are mainly applicable for Tagalog NEs.

For instance, in the phrase “Pasig Mayor Vico Sotto,” the token Pasig would be initially labelled as a location (B-LOC), but would be updated as part of a person entity (B-PER) after matching the full entity using the B-PER dictionary.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Performance evaluation metrics.

PERFORMANCE EVALUATION

DATASETS	TOKEN COUNT	ACCURACY	PRECISION	RECALL	F1-SCORE
TRG	734	0.9959	0.9649	0.9821	0.9735
Ugnayan	1,025	1.0	1.0	1.0	1.0
TweetTaglish	34,424	0.9943	0.9753	0.9981	0.9865
NewsPH-NLI	28,922	0.9730	0.9632	0.9443	0.9537
Pilipino Star Ngayon	5,787	0.9758	0.9852	0.9407	0.9624

Performance evaluation results of PANG-KAT's short-unit tokenization for both unit testing and external validation.

PERFORMANCE EVALUATION

DATASETS	TOKEN COUNT	ACCURACY	PRECISION	RECALL	F1-SCORE
TRG	734	0.9958	0.9473	0.9730	0.9600
Ugnayan	1,025	1.0	1.0	1.0	1.0
TweetTaglish	34,424	0.9931	0.9443	0.9968	0.9698
NewsPH-NLI	28,922	0.9764	0.9034	0.9202	0.9117
Pilipino Star Ngayon	5,787	0.9801	0.9612	0.9036	0.9315

Performance evaluation results of PANG-KAT's longer-unit tokenization for both unit testing and external validation.

PERFORMANCE EVALUATION

PANG-KAT achieved F1 scores exceeding 0.9 on both unit testing and external validation for both short and longer unit tokenization. Given that the F1-score ranges from 0 to 1, with 1 indicating perfect classification performance for the model [22], PANG-KAT F1 scores indicate good performance in accurately tokenizing and classifying Tagalog NEs and MWEs.

DISCUSSION

The deductions on PANG-KAT's F1 score is mainly impacted by its strong reliance on the word patterns in its ruleset; it lacks the ability to recognize deeper contextual meanings and word relationships. Strict word patterns also lead to misclassifications when handling misspellings or incorrect usage of punctuation. Additionally, its dictionary has limited vocabulary and is specifically tailored to the Philippine context.

CONCLUSION

CONCLUSION

The main objective of this study is to develop PANG-KAT to address the lack of language-specific NLP tools for the Tagalog language, particularly a suitable tokenizer. Through rigorous development and testing, the performance evaluation results of PANG-KAT indicated good performance in accurately tokenizing and classifying Tagalog NEs and MWEs. These results affirm the effectiveness of its ruleset and dictionaries in capturing the patterns in Tagalog and Taglish texts.

RECOMMENDATIONS

Further improvements in PANG-KAT's performance can still be achieved by integrating it with additional Tagalog pre-processing modules and expanding its dictionary, all to develop a dedicated Tagalog tokenizer that could serve as a foundation for the development of more advanced Tagalog NLP tools and bridge the gap that impedes its potential NLP advancements.

REFERENCES

[1] M. Herrera, A. Aich, and N. Parde. “TweetTaglish: A Dataset for Investigating Tagalog-English Code-Switching”. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, pp. 2090–2097. 2022. URL: https://aclanthology. org/2022.lrec-1.225/.

[2] J. Gra¨en, M. Bertamini, and M. Volk. “Cutter – a Universal Multilingual Tokenizer”. In: Swiss Text Analytics Conference. Winterthur, 12 June 2018 - 13 June 2018: CEUR-WS, pp. 2090–2097. 2018. URL: https://doi.org/10.5167/uzh-157243.

[3] A. Akkasi, E. Varo˘glu, and N. Dimililer. “ChemTok: A New Rule Based Tokenizer for Chemical Named Entity Recognition”. In: BioMed Research International, pp. 1–9. 2016. URL: https://doi.org/10.1155/2016/4248026.

[4] K. Takaoka et al. “Sudachi: a Japanese Tokenizer for Business”. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European LanguageResources Association (ELRA). 2018. URL: https://aclanthology.org/L18-1355.pdf.

REFERENCES

[5] R. Gordon. “Ethnologue: Languages of the world”. In: SIL International. Dallas, p. 1272. 2005. ISBN:155671159X.

[6] K. Foote. “A Brief History of Natural Language Processing”. In: Dataversity Digital LLC. 2023. URL: https://www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/.

[7] R. E. Roxas, J. M. Imperial, and A. De La Cruz. “Science Mapping of Publications in Natural Language Processing in the Philippines: 2006 to 2020”. In: Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation. Shanghai, China: Association for Computational Linguistics (ACL), pp. 500–509. 2021.
URL: https://aclanthology.org/2021.paclic-1.76/.

[8] A. Aquino and F. de Leon. “Parsing in the absence of related languages: Evaluating low-resource dependency parsers on Tagalog”. In: Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020). Barcelona, Spain (Online): Association for Computational Linguistics, pp. 8–15. 2020. URL: https://aclanthology.org/2020.udw 1.2.pdf.

REFERENCES

[9] R. Friendman. “Tokenization in the Theory of Knowledge”. In: Encyclopedia. Vol. 3. 1, pp. 380–386. 2023. URL: https://doi.org/10.3390/encyclopedia3010024.

[10] V. Singh and B. Saini. “An Effective Tokenization Algorithm for Information Retrieval Systems”. In: Proceedings of the First International Conference on Data Mining (DMIN-2014). Royal Archid, Banglore, India. 2014. URL: https://doi.org/10.5121/csit.2014.4910.

[11] D. Biber. “Representativeness in Corpus Design”. In: Literary and Linguistic Computing. Vol. 8. 4. Associa-tion for Computational Linguistics, pp. 243–257. 1993.

[12] T. McEnery and A. Wilson. “Corpus linguistics”. In: Davies G. (ed.) Information and Communications Technology for Language Teachers (ICT4LT). Slough, Thames Valley University [Online]: De Gruyter Mouton, pp. 1286–1304. 2012.
URL: http://www.ict4lt.org/en/en mod3-4.htm.

[13] D. Biber and J. Jones. “61. Quantitative methods incorpus linguistics”. In: L¨udeling M. Kyt¨o (Ed.), Volume 2: An International Handbook. Berlin, New York: De Gruyter Mouton, pp. 1286–1304. 2009. URL: https://doi.org/10.1515/9783110213881.2.1286.

REFERENCES

[14] S. D. Samson. “A treebank prototype of Tagalog”. In: Bachelor’s thesis. Germany: University of T¨ubingen. 2018.
[15] L. J. Miranda. Developing a Named Entity Recognition Dataset for Tagalog. arXiv: 2311 . 07161 [cs.CL]. 2023.
[16] R. Francisco and A. M. Asis. “Tagalog Dictionary Scraper”. In: GitHub repository. 2023. URL: https://github.com/raymelon/tagalog-dictionary-scraper.
[17] L. Villapando and J. Samaniego. “SApp: Data Visualization and Sentiment Analysis Tool From Twitter Data”. In: 2023. URL: https://lib.ics.uplb.edu.ph/research_paper/1691638468_2023-07_Villapando_Samaniego.pdf
[18] J.L. Fleiss. “Measuring nominal scale agreement among many raters”. In: Psychological Bulletin. Vol. 76. 5, pp. 378–382. 1971.
[19] JSON Editor Online. JSON vs CSV: What is the difference and what should I use? URL: https://jsoneditoronline.org/indepth/compare/json-vs-csv/. 2023.

REFERENCES

[20] J. C. B. Cruz et al. “Exploiting News Article Structure for Automatic Corpus Generation of Entailment Datasets”. In: Pacific Rim International Conference on Artificial Intelligence. 2020. URL: https://doi.org/10.48550/arXiv.2010.11574.

[21] PhilStar. Pilipino Star Ngayon. URL: https://www.philstar.com/pilipino-star-ngayon/. 2025.

[22] C. Bhargava et al. “Depression Detection Using Sentiment Analysis of Tweets”. In: Turkish Journal of Computer and Mathematics Education. Vol. 12. 11. Slough, Thames Valley University [Online]: De Gruyter Mouton, pp. 5411–5418. 2021.

[23] G. Van Rossum and J. L. Drake. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace. 2009. ISBN: 1441412697.

[24] PhilAtlas. Welcome to PhilAtlas. URL: https://www.philatlas.com. 2025.

[25] Department of Budget and Management. Result of Validation of DBM-OCIO for Transparency Seal. URL: https://www.dbm.gov.ph/index.php/result-of-validation-of-dbm-ocio-for-transparency-seal. 2023.

REFERENCES

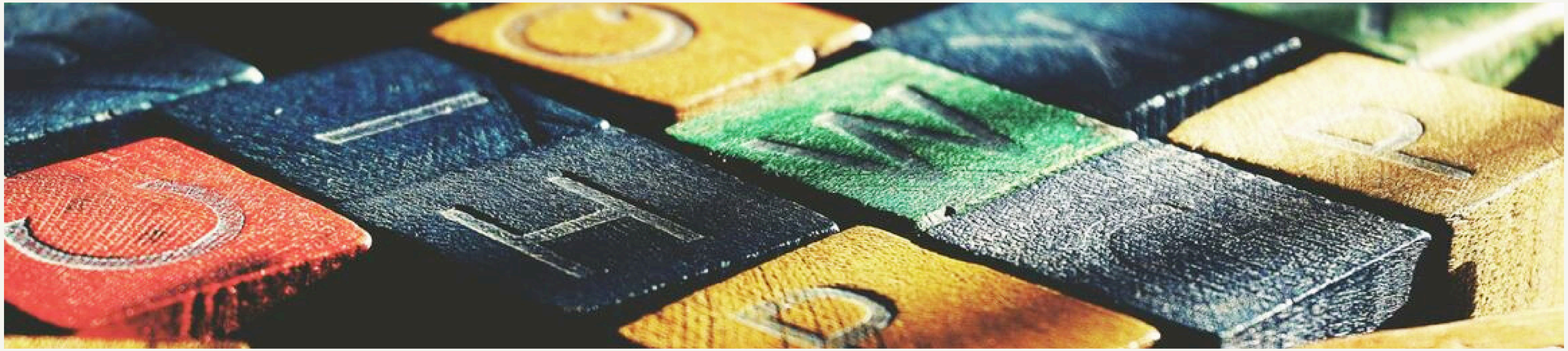
[26] Department of Budget and Management. BESF2024-Acronyms.
URL: https://www.dbm.gov.ph/wp-content/uploads/BESF/BESF2024/ACRONYMS.pdf. 2024.
[27] Philippine Council for NGO Certification. Accredited NGO.
URL: https://pcnc.com.ph/accredited-ngos/. 2025.
[28] P. Remy. “First and Last Name Database”. In: GitHub repository. 2021.
URL: https://github.com/philipperemy/name-dataset.
[29] J. I. Villanueva. “Most Popular Names in the Philippines Dataset”. In: Kaggle. 2023. URL: https://www.kaggle.com/datasets/jorizivannvillanueva/most-popular-names-in-philippines-dataset?resource=download.
[30] Inc. Boo Enterprises. “Filipino Actors / Actresses Celebrities”. In:BooWorld. 2025.
URL: https://boo.world/database/celebrities/filipino-actors-actresses-celebrities.
[31] Central Intelligence Agency. World Leaders - Historical Data.
URL: https://www.cia.gov/resources/world-leaders/historical-data/

REFERENCES

[32] Senate of the Philippines. List of Previous Senators.
URL: https://legacy.senate.gov.ph/senators/senlist.asp. 2022.
[33] V. S. Almario. KWF Manwal sa Masinop Na Pagsulat (Ser. Aklat ng bayan). Komisyon sa Wikang Filipino. 2015. URL: https://kwf.gov.ph/wp-content/uploads/MMP_Full.pdf.
[34] F. De Vos. Essential tagalog grammar: A reference for learners of Tagalog. URL: https://learningtagalog.com/grammar/verbs/pseudoverbs//repeated_pseudoverbs.html.
[35] International Organization for Standardization. ISO 8601 - date and Time Format.
URL: https://www.iso.org/iso-8601-date-and-time-format.html. 2020.

THANK YOU AND HAVE A GOOD DAY!

Justin Louis L. Saavedra and Jaderick P. Pabico



REVIEW OF LITERATURE

Why create a dedicated tokenizer?

NLP tools developed for a specific language suffers in performance when used for other languages. Specificially designed tools for a particular language outperforms because it can better capture the grammatical complexities of the target language [8].

Why tokenizer?

The performance of more advanced NLP tools depends on the quality of tokens generated by the tokenization process, as the quality of tokens is essential for generating a coherent and accurate sequence of tokens that collectively form knowledge [9].

Why do a hybrid rule and dictionary-based approach?

Machine-learning approach for tokenizers requires adequate amount of pre-annotated training data, which the low-resourced Tagalog language lacks [1]. Numerous studies suggest an ideal corpus size of approximately a million tokens for a comprehensive studies on general language [11] [12] [13].