Extracting Filipino Spelling Variants

Nathaniel Oco[1,2], Leif Syliongka[1], Raquel Sison-Buban[2] and Joel Ilao[1]
[1]College of Computer Studies, De La Salle University
[2]College of Liberal Arts, De La Salle University
{nathaniel.oco, leif.syliongka, raquel.sison-buban, joel.ilao}@dlsu.edu.ph

Abstract

We introduce a novel method for extracting Filipino spelling variants from a corpus. As an Austronesian language, Filipino exhibits a high degree of inflectional variability. By leveraging linguistic features, crafting rules, and utilizing a representative dataset, we categorize word pairs into three key groups: those adhering to standard guidelines, deviating forms, and competing norms. Our approach highlights significant overlaps with existing documented spelling variants and underscores the potential for enhanced performance in natural language processing (NLP) tasks. Future research should focus on collaborating with language planning bodies to formulate policy recommendations to streamline standardization efforts.

1 Introduction

The proliferation of spelling variants and errors can hinder the performance of various Natural Language Processing (NLP) tasks, including part-of-speech tagging in German (Scheible et al., 2011), intent classification, slot-filling, and response generation in code-mixed data (Yadav et al., 2022), as well as machine translation and sentiment analysis in Nigerian Pidgin (Lin et al., 2024). Addressing these spelling variants during both in the training and decoding phases can enhance performance across NLP tasks.

In the field of education, analyzing spelling variants is equally important. In the Philippines, a Southeast Asian country with 186 languages according to Ethnologue (Eberhard et al., 2024), several educational tools, such as LanguageTool (Oco and Borra, 2011), Gramatika (Go and Borra, 2016), and Balarila (Ponce et al., 2023), have been developed to correct spelling errors, targeting one of the official Philippine languages—Filipino.

Numerous Filipino spelling variants have been documented in the literature, notably by Zuraw (2006), Ilao et al. (2011), and Gallego (2016). Some examples of these variants are presented in Table 1, where cdiff is the character difference, and Word1 and Word2 have the same meaning.

cdiff	Word1	Word2
d vs. r	madumi	marumi
e vs. i	galeng	galing
o vs. u	kompanya	kumpanya
uw vs. w	kuwento	kwento
iy vs. y	piyano	pyano

Table 1: Spelling variants and examples

One challenge in extracting spelling variants is the occurrence of non-variants or false positives—word pairs that are, in fact, distinct words. Examples are shown in Table 2, with glosses in parentheses. This paper aims to address this issue by proposing a methodology for extracting spelling variants from a corpus, utilizing linguistic features and carefully crafted rules. Our contributions can be summarized as follows:

- We identified various linguistic features and created rules to extract word pairs that are spelling variants;
- We conducted experiments on a monolingual corpus of Filipino texts; and
- We categorized word pairs into three distinct types based on their alignment with existing guidelines.

Our approach has implications for both educational tools and larger NLP applications that rely on accurate word forms for efficient processing.

Filipino language

The focus of this study is the Filipino language, which is characterized by free word order and a high degree of inflection. Beyond education, the extraction of spelling variants plays a critical role

cdiff	Word1	Word2
d vs. r	madikit (sticky)	marikit (pretty)
e vs. i	pare (buddy)	pari (priest)
o vs. u	opo (yes)	upo (eggplant)
uw vs. w	pauwi (go home)	pawi (erase)
iy vs. y	paiyak (to cry)	payak (simple)

Table 2: Example of non-variants

in language standardization. According to a report by National Geographic (Rymer, 2012), one language dies every 14 days, and nearly half of the approximately 7,000 spoken languages worldwide are expected to disappear within the next century (Anderson, 2010). Documenting and compiling dictionaries is an essential step in preserving endangered languages, while standardization ensures consistency and usability in lexicographic work.

In the Philippines, data from Ethnologue (Eberhard et al., 2024) reveals 186 documented languages, making the country a linguistic treasure trove. Of these, nine are non-indigenous, 175 are indigenous, and two have already become extinct. These statistics underscore the urgent need for a comprehensive databank of Philippine languages and highlight the crucial importance of standardization in preserving this rich linguistic heritage.

The Komisyon sa Wikang Filipino (KWF), also known as the Commission on the Filipino Language (CFL), was established under the 1987 Constitution of the Philippines [1]. It serves as the official regulatory body responsible for the development, preservation, and promotion of Filipino and other local Philippine languages [2]. The Philippine orthography has evolved from 20 letters in 1940 to 28 letters in 1987:

- 1940: a, b, k, d, e, g, h, i, l, m, n, ng, o, p, r, s, t, u, w, y
- 1987: addition of eight letters {c, f, j, ñ, q, v, x, z}

Ten years ago, the KWF released the 2014 edition of the National Orthography (sa Wikang Filipino, 2014), which provides guidelines for writing the Filipino language and was used to match the results of our experiments.

Extracting spelling variants

Linguistic features

To extract linguistic features, word unigram models and character n-gram profiles of a given corpus need to be generated. Preprocessing involves tokenization and true-casing. The features we considered are:

- edit distance, a measure of how different two strings (or sequences of characters) are from one another, which is defined as the minimum number of operations (character insertion, deletion, or replacement) needed to transform one string into another (Levenshtein, 1966);
- string length;
- cdiff or character difference to show additions and deletions of characters;
- cdiff index, the index where the cdiff occurred;
- cdiff position (beginning, middle, ending of a word);
- character n-grams, with a minimum value of 3 (trigram) and a maximum value of 4 (4-gram); and
- generalized character n-grams, where consonants and vowels are generalized.

Edit distance has been widely used in cognate and spelling variants detection (Babych, 2016; Messner and Lippincott, 2024; Barteld, 2017; Laarmann-Quante et al., 2022) but there is limited attempt in the past to utilize character n-grams in detecting Filipino spelling variants.

Rule creation

Machine learning is used to identify significant features by constructing a feature set and labeling word pairs as either spelling variants or non-variants. Attribute evaluators (Hall and Smith, 1999), particularly rankers, guide the rule creation process. The results highlight the features that effectively identify spelling variants. Previous studies (Ilao et al., 2011; Gallego, 2016) relied on manual selection. To the best of our knowledge, this is the first attempt to apply a machine learning approach to determine Filipino spelling variants. Once the rules are created, they can be transformed

[1] Article XIV, Section 6

[2] https://kwf.gov.ph/mandato/

into regular expressions to efficiently match patterns.

Experimental setup

Data and tools

The corpus used in this study is the August 20 snapshot[3] of the Tagalog Wikipedia (Contributors, 2024). Wikipedia is available in different languages and the Tagalog Wikipedia serves as a representation of the Filipino language [4]. The raw corpus contains 11 million words and 68 million characters. We employed SRILM (Stolcke, 2002) and Apache Tika (Mattmann and Zitting, 2011) to generate word unigram models and character n-gram profiles of the corpus, respectively. Wdiff (Pinard, 1992) was used to identify character differences, while the Waikato Environment for Knowledge Analysis (Weka) (Witten et al., 2011) was used for attribute evaluation. We also utilized Notepad++[5] to convert the rules to regular expressions, enabling the efficient extraction of word pairs.

Additionally, a spreadsheet application and several custom-developed programs were used to automate the population of the feature set, a sample of which is shown in Table 3. The complete feature set includes 4-grams, though these are omitted from the table for clarity. Various n-gram configurations were explored, including cases where the cdiff index is the first letter (n-gram1), second letter (n-gram2), and so on (n-gram3 and n-gram4). For example, if the cdiff corresponding to [-a-]+i+, and with ’t’ and ’n’ as the characters to the left and right, respectively, the 3-grams2 are "tan" and "tin." The notation "gen" (e.g., 3-gram1gen) stands for "generalized," where consonants are replaced with ’C’ and vowels with ’V’. The "Class" refers to the label, indicating whether a pair is a variant or non-variant.

Limitations

Due to the multilingual nature of the Philippines, code-switching is inevitable. English words were excluded. Additionally, due to the number of variables involved, the method is limited to an edit distance of 1. Proper nouns and variations at the morphological level, including reduplication, were also excluded.

Feature	Example
word1	aabutan
word2	aabutin
edit Distance	1
string length1	7
string length2	7
cdiff	[-a-]+i+
cdiff index	6
cdiff position	middle
3-gram1 word1	an_
3-gram1 word2	in_
3-gram1gen word1	aC_
3-gram1gen word2	iC_
3-gram2 word1	tan
3-gram2 word2	tin
3-gram2gen word1	CaC
3-gram2gen word2	CiC
3-gram3 word1	an_
3-gram3 word2	in_
3-gram3gen word1	aC_
3-gram3gen word2	iC_
Class	non-variant

Table 3: Sample feature set

cdiff	Word1	Word2
d vs. r	madami	marami
e vs. i	aatakehin	aatakihin
o vs. u	abogado	abugado
uw vs. w	kuwintas	kwintas
iy vs. y	piyansa	pyansa

Table 4: Spelling variants reported in other works

Results and discussion

Spelling variants identified

We were also able to extract spelling variants detected in earlier studies. These spelling variants are shown in Table 4. We noted that only using cdiff would also result to false positives if English words are also extracted (e.g., robber vs. rubber and polling vs. pulling for o vs. u). The list of rules that yielded 100% precision rate for Filipino word pairs, totaling four, are in Table 5, where ’C’ is for consonant and ’V’ is for vowel. These four rules cover 807 word pairs and the manually-validated data is publicly available online[6]. Exploring various n-gram configurations as part of the feature set proved advantageous.

[3] https://dumps.wikimedia.org/tlwiki/20240820/tlwiki-20240820-pages-articles.xml.bz2

[4] https://en.wikipedia.org/wiki/Tagalog_Wikipedia

[5] https://notepad-plus-plus.org/

[6] Public data: https://forms.gle/9gvvu2KYfvAF2wR86

cdiff	Example
ehVC vs. ihVC	doblehin vs. doblihin
omC vs. umC	kompanya vs. kumpanya
CuwV vs. CwV	lenggwahe vs. lengguwahe
CiyV vs CyV	ahensiya vs. ahensya

Table 5: Rules with 100% precision

cdiff	Abecedario	Modern orthography
c vs. k	acalain	akalain
o vs. w	dinadalao	dinadalaw
i vs. y	baitang	baytang
v vs. b	automovil	automobil

Table 6: Abecedario and the modern orthography

Abecedario

Our approach was also able to detect word pairs that reflect both the Abecedario and the modern orthography. The Abecedario is the alphabet used in the early Spanish-influenced orthography of Filipino during the Spanish colonial period. It is derived from the Spanish alphabet and was widely used before the introduction of modernized and standardized forms of orthography. Some examples are provided in Table 6, highlighting the potential for conducting culturomics studies.

Alignment with existing guidelines

For each word pair, we counted the number of occurrences in the corpus and converted these counts into percentages, with the total for both words always adding up to 100%. We observed that certain word pairs, where one form appears 40% of the time or less, do not align with the 2014 edition of the National Orthography. An example under omC vs. umC is "kumpleto" ("complete" in English) with 86% compared to "kompleto" (14%), which is the word listed in the KWF dictionary. Additionally, we identified competing word forms, which we defined as having frequencies between 41% and 60%. Examples of competing forms are shown in Table 7. The percentages are enclosed in parenthesis.

We categorize word pairs into three:

1. those adhering to existing guidelines (61 to 100%);
2. competing norms (41 to 60%) and
3. deviating forms (up to 40%).

Word1 (%)	Word2 (%)
komplikado (42%)	kumplikado (58%)
kompanya (51%)	kumpanya (49%)
kompirmasyon (53%)	kumpirmasyon (47%)
pinupuwersa (50%)	pinupwersa (47%)
lisensiya (48%)	lisensya (52%)

Table 7: Examples of competing norms

In 2018, several years after the release of the 2014 edition, a contest hosted by the Komisyon sa Wikang Filipino (KWF) revealed students' deficiencies in Filipino orthography (De Guzman, CG, 2018). Out of a perfect score of 100, the first place only got a score of 65. These findings underscore the need for technological tools that comply with KWF guidelines such as spell checkers that are freely available and convenient to use.

Cosine similarity

We conducted additional experiments to determine whether the word pairs share semantic meaning. Using Word2Vec (Mikolov et al., 2013), we applied a continuous bag-of-words model (Rong, 2014) with a word window of 5 to compute cosine similarity values. This is inspired by an earlier work which looked at English words (Jatnika et al., 2019). Our results show high similarity values among competing norms with high frequency counts. Low similarity values were noted in Wikipedia articles that appear to be machine translated.

Conclusion

We developed an effective method for extracting spelling variants from a corpus. Through experiments with the Tagalog Wikipedia, we successfully extracted features and created rules using machine learning. As a next step, our findings can be integrated into widely-used Filipino spelling, style, and grammar checking tools to enhance their accuracy and functionality. Furthermore, collaboration with institutions such as the Komisyon sa Wikang Filipino (KWF) could facilitate the consistent application of standardized Filipino spelling across various platforms, promoting linguistic uniformity while supporting language education and preservation. Legitimate variants, including those classified as competing norms and deviating forms, should receive special attention in Filipino language education.

References

Stephen Anderson. 2010. How many languages are there in the world? Linguistic Society of America.

Bogdan Babych. 2016. Graphonological Levenshtein edit distance: Application for automated cognate identification. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 115–128.

Fabian Barteld. 2017. Detecting spelling variants in non-standard texts. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 11–22, Valencia, Spain. Association for Computational Linguistics.

Contributors. 2024. Tagalog wikipedia, the free encyclopedia. https://tl.wikipedia.org/. [Online; accessed 9-September-2024].

De Guzman, CG. 2018. Contest Result Shows Students' Deficiency in Filipino Orthography. https://www.ptvnews.ph/contest-result-shows-students-deficiency-in-filipino-orthography/. [Online; accessed 9-September-2024].

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World*, twenty-seventh edition. SIL International, Dallas, Texas. Online version: `http://www.ethnologue.com`.

Maria Kristina Gallego. 2016. Isang pagsusuri sa korpus ukol sa pagbabago ng wikang filipino, 1923-2013. *Philippine Social Sciences Review*, 68(1):71–101.

Matthew Phillip Go and Allan Borra. 2016. Developing an unsupervised grammar checker for Filipino using hybrid n-grams as grammar rules. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*, pages 105–113, Seoul, South Korea.

Mark A. Hall and Lloyd A. Smith. 1999. Feature subset selection: A correlation based filter approach. In *Proceedings of the 1999 International Conference on Neural Information Processing and Intelligent Information Systems*, pages 855–858, Perth, Australia.

Joel Ilao, Rowena Cristina Guevara, Virgilio Llenaresas, Eilene Antoinette Narvaez, and Jovy Peregrino. 2011. Bantay-wika: towards a better understanding of the dynamics of Filipino culture and linguistic change. In *Proceedings of the 9th Workshop on Asian Language Resources*, pages 10–17, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Derry Jatnika, Moch Arif Bijaksana, and Arie Ardiyanti Suryani. 2019. Word2vec model analysis for semantic similarities in english words. *Procedia Computer Science*, 157:160–167. The 4th International Conference on Computer Science and Computational Intelligence (ICCSCI 2019) : Enabling Collaboration to Escalate Impact of Research Results for Society.

Ronja Laarmann-Quante, Leska Schwarz, Andrea Horbach, and Torsten Zesch. 2022. 'meet me at the ribary' – acceptability of spelling variants in free-text answers to listening comprehension prompts. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 173–182, Seattle, Washington. Association for Computational Linguistics.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710.

Pin-Jie Lin, Merel Scholman, Muhammed Saeed, and Vera Demberg. 2024. Modeling orthographic variation improves nlp performance for nigerian pidgin. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 11510–11522.

Chris Mattmann and Jukka Zitting. 2011. *Tika in Action*. Manning Publications Co., Greenwich, CT, USA.

Craig Messner and Thomas Lippincott. 2024. Pairing orthographically variant literary words to standard equivalents using neural edit distance models. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 264–269, St. Julians, Malta. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR) 2013*.

Nathaniel Oco and Allan Borra. 2011. A grammar checker for Tagalog using LanguageTool. In *Proceedings of the 9th Workshop on Asian Language Resources*, pages 2–9.

Francois Pinard. 1992. *GNU wdiff manual*. Free Software Foundation.

Andre Dominic H. Ponce, Joshue Salvador A. Jadie, Paolo Edni Andryn Espiritu, and Charibeth Cheng. 2023. Balarila: Deep learning for semantic grammar error correction in low-resource settings. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 21–29, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Xin Rong. 2014. word2vec parameter learning explained. *ArXiv*, abs/1411.2738.

Russ Rymer. 2012. Vanishing voices. National Geographic.

Komisyon sa Wikang Filipino. 2014. *Ortograpiyang Pambansa*. Komisyon sa Wikang Filipino, Manila.

Silke Scheible, Richard J. Whitt, Martin Durrell, and Paul Bennett. 2011. Evaluating an 'off-the-shelf' pos-tagger on early modern german text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 19–11522.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Seventh International Conference on Spoken Language Processing*. ISCA.

Ian H. Witten, Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edition. Morgan Kaufmann.

Krishna Yadav, Md Akhtar, and Tanmoy Chakraborty. 2022. Normalization of spelling variations in code-mixed data. In *Proceedings of the 19th International Conference on Natural Language Processing*, pages 269–279.

Kie Zuraw. 2006. Using the web as a phonological corpus: A case study from Tagalog. In *Proceedings of the 2nd International Workshop on Web as Corpus*.