IEEE Region 10 TENCON 2026 IEEE Indonesia Section

Orthographic Nativization as Rule-Based Rewrite Cascades

Erin Gabrielle Chua
Zhean Robby Ganituen
Justin Ethan Ching
Jaztin Jacob Jimenez
Nathaniel Oco

De La Salle University, Manila, Philippines

The Filipino NLP Challenge

Orthographic Nativization

Borrowed words modify spelling to obey recipient language phonotactics and orthography:

- "computer" –"kompyuter"
- "basketball" – "basketbol"
- "police" – "pulis"

Downstream NLP Impact

Accurate algorithmic modeling of these mutations directly powers key language infrastructure:

- **Text Normalization**: Standardizing social media & mixed-language feeds.
- **Lexical Disambiguation**: Aiding spell-checkers and speech recognition.
- **Corpus Expansion**: Clean indexing of low-to-medium resource vocabularies.

Orthography Matters Too!

Computational loanword research has overwhelmingly concentrated on acoustic phonology and morphological inflection. In contrast, **orthographic adaptation remains severely understudied** despite text-based systems requiring exact spelling harmonization.

This work closes that critical computational blindspot by evaluating deterministic spelling transformations in formal written text.

Filipino Orthography & Linguistic Challenges

Vowel Ambiguity

Filipino has 5 vowel letters ('a', 'e', 'i', 'o', 'u'). English's larger inventory causes many-to-one mappings (e.g., /æ/ and /u:/ both map to 'a').

Schwa Reduction

English unstressed vowels reduce to schwa /ə/, unencoded in orthography, requiring pronunciation data for resolution.

Vowel Interchangeability

Vowels ('e', 'i', 'y') and ('o', 'u') are often treated as interchangeable in many lexical contexts.

Filipino Orthography & Linguistic Challenges

Consonant Mappings

English sounds lacking native letters (e.g., /f/, /tʃ/, /dʒ/) have fixed prescriptive forms ('s'/'sy', 'ts', 'dy'). Some nativize based on adjacent graphemes.

Syllable Constraints

Filipino favors CV or CVC structures. Illegal consonant clusters require epenthetic vowel insertion.

Proposed System Pipeline & Priority Rewrite Cascade

Stage 1
Graphemic Tokenization

Segment Input

Segments English input text into distinct structural units.

Multi-Letter Recognition

Recognizes complex graphemic structures: **Digraphs, Trigraphs, Tetragraphs, etc.**

Stage 2
Rule-Based Mapping

Priority Rewrite Cascade

Executes ordered rules sequentially for precise grapheme transformations.

Rule Hierarchy:

1 Context-sensitive rules

2 Phonetic rules

3 Context-free mappings

Stage 3
Phonetic Resolution

G2P Lookup

Leverages G2P lookup dictionary and heuristics.

Vowel Disambiguation

Resolves ambiguous, context-dependent vowel sounds based strictly on target pronunciation metrics.

Stage 4
Output Normalization

Cleanup

Cleans up residual structural artifacts from intermediate stages.

Phonotactic Cluster Repair

Applies natural cluster repair rules (e.g., epenthetic 'o' insertion for valid phonotactics).

System Architecture Advantage:

Rules are hand-authored from prescriptive guidelines and linguistic descriptions, requiring no parallel training corpora.

Experimental Setup & Gold Standard

Gold Standard

2,319

English-Filipino Loanword Pairs

- **1,858 pairs:** common everyday speech
- **461 pairs:** drug names

Annotators

3 Native

Mesolectal Speakers

Independently annotated the list based on **KWF guidelines** and plausible forms under the **5-vowel system**.

Agreement

92%

Majority-Match

Inter-annotator agreement before adjudication.

Disagreements resolved via discussion between annotators.

Evaluation Metric

CER

Character Error Rate

Calculated as **Levenshtein edit distance** normalized by reference character count.

Key Experimental Results & Baseline Comparisons

PROPOSED SYSTEM (IPA-GUIDED)

5.14%
Character Error Rate

Strict 5-Vowel Evaluation
Evaluated against exact 5-vowel representations without equivalence relaxation.

4.05%
Character Error Rate

3-Vowel Evaluation
Relaxed evaluation combining equivalence classes for **'e,i,y'** and **'o,u'**.

BASELINE COMPARISONS (5-VOWEL CER)

Spelling-only Ablation
Ablation model without IPA-guided components
17.73%

Regex-based System
Rule-based simple regex substitution baseline
20.38%

Large Language Model
Prompting-based large language model baseline
15.54%

Key Conclusion:
The rewrite cascade significantly outperforms prompting-based approaches and simple regex substitutions, achieving up to a **15.24% absolute reduction in CER** over standard regex baselines.

Error Analysis & Dominant Failure Modes

Vowel Ambiguity

Unstressed Vowel Ambiguity

Largest error source. English schwa lacks a direct Filipino equivalent, forcing arbitrary mappings (e.g., 'e'/'i').

- **Accounts for >133 vowel substitutions under 3-vowel evaluation.**

Glide Insertion

Incomplete Glide Insertion

Filipino requires glides /w/ and /j/ to break vowel hiatus.

- **Missed 42 'w'-insertion & 40 'y'-insertion sites (e.g., 'tiara' instead of 'tiyara').**

G2P Error Propagation

G2P Propagation Errors

Incorrect upstream IPA transcriptions directly affect downstream output.

- **40 edits trace to G2P errors, concentrated in long compound words & pharmaceutical terms.**

Conclusion & Future Directions

Key Conclusion

Prescriptive linguistic rules can be successfully operationalized into an effective **rewrite-cascade pipeline** for orthographic nativization **without requiring large parallel corpora**.

Future Work & Research Directions

Stress Resolution

Lexical Stress

Incorporate lexical stress information directly into the pipeline to improve schwa resolution accuracy.

Hybrid Models

Rule + Data

Explore hybrid rule-and-data approaches to achieve further performance gains on complex edge cases.

Scalability

Expansion & Domain

Extend the architecture to other language pairs and evaluate robustness on informal register text.

Thank You!

For contact information and more info, please visit the project's webpage.