DOI: https://10.83080/rejost.vol6no10.366

Researchers Journal of Science and Technology (REJOST)

A Computational Analysis of Lexical and Form-Level Variation between the Lutenga and Lupakoyo Varieties of Lusoga

Azizi Wasike[1, 2*], Ismail Kamukama[1, 3], Adeleke Raheem Ajiboye[1], Awadallah M. Ahmed[1], Muriithi Irene Wanjiru[1]

Abstract

Differences between closely related language varieties are not always apparent from general measures of similarity, particularly when they involve alternative words or recurring changes in written forms. This study examined lexical and written-form variation between Lutenga and Lupakoyo, two regional varieties of Lusoga spoken in eastern Uganda. Using 14,597 aligned English–Lutenga–Lupakoyo sentence triplets, the analysis combined sequence-based word alignment, character-similarity analysis, character-edit analysis, contextual screening, and bidirectional correspondence-consistency measures. It identified 5,644 distinct candidate word-form correspondences, of which 422 occurred at least five times. Among these recurrent pairs, 41 had character-similarity scores below 0.40. Examination of the corresponding translation contexts suggested that eight of these pairs could represent alternative lexical choices, although their linguistic status was not independently established. Most extracted correspondences (75.6%) involved closely resembling written forms, although this resemblance does not necessarily imply a shared linguistic structure. The analysis also identified 52 character-edit families. Insertion of o was the most frequent, occurring 2,699 times across ten distinct word pairs, whereas deletion of g occurred 526 times across 31 pairs. This difference illustrates how a transformation may occur frequently without being widely distributed across word forms. Of the 422 recurrent correspondences, 369 (87.4%) were mutual top counterparts, while 266 (63.0%) achieved at least 80% bidirectional consistency.These measures indicate recurring and relatively consistent associations within the extracted inventory, but do not independently establish that the correspondences represent systematic linguistic differences between the two varieties. The study provides a corpus-based account of these associations, offering a basis for investigating lexical differentiation and written-form variation, and for identifying patterns that warrant further linguistic validation.

✉ Azizi Wasike: a.wasike@muni.ac.ug

[1]Department of Computer Science, Kampala International University, Kampala, Uganda

[2]Department of Computer and Information Science, Muni University, Arua, Uganda

[3]Department of Data Science, Networks & Artificial Intelligence, Kyambogo University, Kampala, Uganda

Keywords: Lusoga; Lutenga; Lupakoyo; Lexical variation; Character-level analysis; Correspondence consistency

@ REJOST 2026

Article history
Received: September 10, 2026
Received in revised form: November 8, 2026
Accepted for publication: November 11, 2026
Available online: November 21, 2026

1.0 Introduction

1.1 Background and Research Context

Closely related language varieties often share much of their vocabulary and grammatical structure, yet differ in how particular meanings are expressed. Some differences involve the choice of words, while others appear in the written forms of related expressions. These distinctions can be difficult to recognise through general measures of linguistic similarity, especially when two varieties resemble each other in most respects but differ in particular contexts. Examining corresponding words and expressions can therefore reveal patterns that broader comparisons might overlook.

Such differences have attracted growing attention in computational linguistics. Joshi et al. (2025), in their survey of natural language processing for dialects, discuss the challenges that linguistic variation presents for computational methods and language resources. Shim and Plank (2025) approach the question from another perspective, showing through their investigation of Italian dialects that variation does not always follow clearly defined dialect boundaries. Combined, these studies suggest that understanding closely related varieties requires attention to the particular ways in which they differ, rather than assuming that each variety is internally uniform.

Interpreting these differences, however, is not always straightforward. Two expressions may differ because speakers use alternative words, because related words are written differently, or because their grammatical structures vary. Lexical variation concerns differences in word choice, orthographic variation concerns written representation, and morphological variation involves differences in word structure. In translated texts, another possibility arises: translators may choose different expressions even when working from the same source sentence. A difference identified through computational comparison may therefore have several explanations, and its linguistic significance cannot always be determined from the written forms alone.

These questions are particularly relevant to Lusoga, a Bantu language spoken predominantly in the Busoga sub-region of eastern Uganda. Among its regional varieties are Lutenga and Lupakoyo, associated with the southern and northern dialect clusters, respectively. Previous linguistic and corpus-based research has contributed to understanding Lusoga, but the availability of aligned English–Lutenga–Lupakoyo translations offers an opportunity to examine relationships between the two varieties more closely. Because the translations share the same English source sentences, corresponding expressions can be compared within similar contexts to identify recurring differences in vocabulary and written forms. Although some of these differences may reflect individual translation choices, examining their recurrence and consistency provides a useful starting point for understanding the variation represented in the corpus.

1.2 Linguistic and Computational Perspectives on Language-Variety Variation

Computational approaches have made it possible to examine linguistic variation without relying entirely on predefined dialect features. Kuparinen and Scherrer (2024), for example, investigated dialectal variation using interview transcriptions from Finnish, Norwegian and Swiss German. They applied topic modelling to different representations of the texts, including complete words, character sequences and automatically segmented forms. Their findings showed that the choice of representation influences the patterns that computational methods can identify. This is particularly relevant where detailed linguistic annotation is

unavailable, as differences between language varieties can still be explored through the forms recorded in a corpus.

Not every difference between comparable expressions, however, can be attributed to dialect. Shadrova (2025), examining lexical variability in task-based corpora, showed that word choices may differ even under similar communicative conditions. Such variability complicates the interpretation of recurring correspondences, particularly when the data come from translations produced by different individuals. A word may appear more frequently in one variety without necessarily being restricted to that variety. Its occurrence may also reflect the translator's preferences, the surrounding grammatical context or the alternatives available for expressing a particular meaning.

These difficulties are especially relevant when comparing written forms in Bantu languages. Corresponding expressions may involve different lexical items, grammatical elements attached to word stems or alternative written representations of related forms. Edelsten et al. (2024) examine morphosyntactic variation in Bantu languages through explicitly defined grammatical parameters, illustrating why differences that appear minor at the character level may involve more complex linguistic relationships. Creissels (2025) provides a broader account of structural variation within the Bantu language family, while Kasonde (2025), in a study of Icibemba orthography, draws attention to the difficulties that variable spelling practices present for digital language representation. These studies show why character-level differences are useful evidence for comparison but cannot, on their own, establish whether the underlying variation is lexical, orthographic or morphological.

Computational studies have approached these questions from different directions. Xie et al. (2024) investigated interpretable dialect classification using Mandarin, Italian and Low Saxon data, examining how lexical features could help explain distinctions between language varieties. Their approach makes it possible to identify words associated with particular dialects, although identifying distinctive vocabulary is not the same as establishing relationships between corresponding expressions. A word may help distinguish one variety from another without revealing which form expresses a comparable meaning in a related variety. Aligned translations offer another way of examining these relationships because expressions associated with the same source sentence can be compared directly.

Written-form variation also has implications for language technologies. Lin et al. (2024) examined orthographic alternatives in Nigerian Pidgin and found that incorporating such variation into training data improved performance in the sentiment analysis and machine translation tasks they investigated. Their findings demonstrate that differences in written representation can affect computational processing, although the extent and nature of these effects may vary across languages and applications.

The existing approaches therefore provide different but complementary ways of investigating language variation. Dialect classification helps identify features associated with particular varieties, while word-level and character-level comparisons reveal differences between individual forms. For the present study, the interest lies in the relationships between corresponding Lutenga and Lupakoyo words: which forms recur together, how their written representations differ and how consistently those associations appear across aligned translations. Examining these relationships provides a more detailed account of the variation recorded in the corpus without assuming that every extracted correspondence represents an established linguistic distinction.

1.3 Lusoga and Its Regional Varieties

Lusoga is a Bantu language spoken predominantly in the Busoga sub-region of eastern Uganda. Although commonly discussed as a single language, it encompasses regional varieties that share much of their linguistic structure while exhibiting differences in vocabulary, pronunciation and grammatical usage.Lutenga and Lupakoyo represent two broadly recognised dialect groupings, traditionally associated with southern and northern Busoga, respectively, although their geographical distribution is not confined to strict regional boundaries (Isiko, 2018). Their close relationship makes them particularly relevant for examining variation that may be difficult to capture through general descriptions of the language.

Existing scholarship has provided an important foundation for understanding Lusoga. Nabirye (2016) contributed a corpus-based grammatical description of the language, while Nabirye et al. (2016) examined the phonetic characteristics of Lutenga. These studies offer insights into Lusoga's linguistic structure and the characteristics of one of its regional varieties. However, describing the grammatical or phonetic properties of a language is different from systematically examining how corresponding words and expressions vary between two regional varieties.

Corpus-based research has also contributed to the documentation and computational study of Lusoga. de Schryver and Nabirye (2018) examined the development of Lusoga corpus resources, demonstrating the value of systematically collected language data for linguistic investigation. Their subsequent work on neology detection (de Schryver & Nabirye, 2022) further illustrates how corpus analysis can be used to investigate patterns of lexical usage and change. These contributions show the potential of computational approaches for studying Lusoga, particularly where linguistic resources remain limited.

The present study builds on this foundation by examining aligned translations in Lutenga and Lupakoyo. The corpus contains English source sentences accompanied by translations into both varieties, allowing corresponding expressions to be compared within a shared source context. This arrangement makes it possible to investigate recurring lexical correspondences and differences in written forms, rather than considering each variety independently. At the same time, the translations represent particular choices made by their contributors, and the patterns observed in them cannot automatically be taken as evidence that a word or form belongs exclusively to one variety. The analysis therefore concentrates on the recurrence, distribution and consistency of correspondences within the available corpus.

1.4 Research Problem and Gap

Research on linguistic variation has increasingly benefited from computational methods that identify differences across language varieties. Much of this work has focused on distinguishing dialects, measuring their overall similarity or identifying linguistic features associated with particular varieties. While these approaches have improved our understanding of dialectal variation, they do not always show how individual words correspond across closely related varieties. Computational alignment has also been used to examine lexical relationships between closely related varieties. Min et al. (2024), for example, aligned parallel texts involving Standard Malay and two Malay dialects to investigate their lexical similarities and differences. However, identifying such associations does not, on its own, explain how they vary in frequency, written form or consistency between particular regional varieties. Such relationships are particularly important when differences involve alternative lexical choices or recurring changes in written forms that broader measures of similarity may overlook.

For Lusoga, existing studies have examined grammatical structure (Nabirye, 2016), the phonetic characteristics of Lutenga (Nabirye et al., 2016), corpus development (de Schryver & Nabirye, 2018) and lexical development through neology detection (de Schryver & Nabirye, 2022). Although these studies have contributed to our understanding of Lusoga, direct computational comparisons of corresponding Lutenga and Lupakoyo word forms using aligned translations remain limited in the literature reviewed. In particular, the available research offers little evidence of how frequently particular word-form correspondences recur, how their written representations differ and how consistently the forms are associated across comparable translation contexts.

Examining these relationships requires care because differences between corresponding translations may arise from several sources. Alternative words may reflect lexical variation, while differences in spelling or word structure may involve orthographic or morphological variation. Other differences may result from the choices made by translators working from the same English sentences. The present study therefore examines recurring word-form correspondences, character-level differences and correspondence consistency within an aligned English–Lutenga–Lupakoyo corpus. The analysis thus describes the patterns found in the translations without assuming that every extracted difference represents an established dialect feature.

1.5 Research Aim, Questions and Contributions

The study aimed to examine recurrent lexical and written-form variation between Lutenga and Lupakoyo using aligned translations of the same English source sentences. Rather than treating the two varieties as separate collections of words, the analysis focused on correspondences between their forms, the differences observed in their written representations, and the extent to which these relationships recurred across the corpus.

Three research questions guided the investigation:

1. What recurrent lexical correspondences characterise Lutenga and Lupakoyo when translations of the same English source sentences are compared?
2. What recurrent written-form patterns characterise differences between corresponding Lutenga and Lupakoyo word forms?
3. How frequently do the identified correspondences recur, and how consistently are the forms associated with one another across the corpus?

The study contributes new empirical evidence on lexical correspondences and written-form variation between Lutenga and Lupakoyo, two closely related varieties of Lusoga that have received limited direct computational comparison. Its contribution lies not in developing a new alignment or similarity algorithm, but in providing a corpus-based account of how individual word-form correspondences recur, differ in written representation and vary in directional consistency across the two varieties. By examining these relationships together, the study distinguishes frequently observed associations from those that are more consistently paired or distributed across a wider range of word forms. The analysis distinguishes the frequency of recurring character transformations from their distribution across distinct word pairs, allowing frequently observed differences to be considered alongside the range of forms in which they occur. It also examines correspondence consistency in both directions, providing a more detailed account of the relationships between extracted word forms than occurrence counts alone. Robustness checks further examine how the identified correspondences vary with different frequency thresholds, character-similarity cut-offs and alignment methods. These findings offer an empirical basis for investigating variation represented in the corpus

and may inform future work on variety-sensitive Lusoga language resources. They do not, however, independently establish dialect-exclusive vocabulary or productive linguistic rules.

2.0 Materials and Methods

2.1 Research Design and Analytical Approach

The study adopted a quantitative, exploratory corpus-based approach to examine lexical and written-form variation between Lutenga and Lupakoyo. The analysis used aligned English–Lutenga–Lupakoyo sentence triplets, allowing translations of the same English source sentence to be compared across the two varieties. This provided a common source context for identifying corresponding words and examining differences in their written forms.

The analysis proceeded from word-level correspondence extraction to character-level comparison and the assessment of recurrence and consistency. Sequence-based alignment was used to identify candidate correspondences between Lutenga and Lupakoyo word forms, while character-similarity measures helped distinguish closely resembling forms from those with more substantial written differences. Recurrent correspondences were then examined to identify possible lexical alternatives and recurring character transformations. Frequency counts, the distribution of transformations across distinct word pairs, and bidirectional correspondence-consistency measures were used to describe the patterns represented in the corpus.

The approach was exploratory rather than a formal linguistic classification of dialect features. Word alignments and character comparisons provided evidence of relationships between forms, but they did not independently establish whether a particular difference arose from lexical choice, orthography, morphology or translation preferences. The analysis therefore treated the extracted pairs as candidate correspondences and interpreted their recurrence and consistency within the limits of the available translated data.

2.2 Corpus Source and Preparation

The study drew on an aligned English–Lutenga–Lupakoyo corpus developed as part of a broader research project on English–Lusoga machine translation. The English source sentences originated from the 41,070 English–Luganda sentence pairs reported by Kimera et al. (2022), although only the English component was reused. The translations were produced by 18 Lusoga A-level students, comprising nine native Lutenga speakers and nine native Lupakoyo speakers. Each translator was assigned 1,669 English sentences, giving 15,021 translations per variety. The two groups worked independently, with each participant translating exclusively into their native variety. All translators followed common translation guidelines and were informed of the research objective. The initial corpus contained 15,021 sentence triplets, which were reduced to 14,597 after excluding records with incomplete or unusable target translations and removing repeated entries. Each retained triplet comprised an English sentence and its corresponding Lutenga and Lupakoyo translations, with the two varieties maintained separately to allow direct comparison without treating either as the reference standard.

The analysis covered all 14,597 retained triplets, using the shared English sentences to provide comparable contexts for examining lexical and written-form differences between the two varieties. Although this arrangement helped limit differences arising from unrelated source content, it could not eliminate variation associated with individual translation choices or grammatical constructions. The training, validation and test partitions prepared for the broader machine translation project were not used to restrict correspondence extraction in this study. Instead, all 14,597 retained triplets were analysed together, with the resulting correspondences describing variation across the

complete corpus rather than providing an independent assessment on held-out data.

2.3 Word-Level Correspondence Extraction

Word-level correspondences between Lutenga and Lupakoyo were extracted using a sequence-based comparison procedure implemented in Python. Before alignment, the translations were converted to lowercase, punctuation at word boundaries was removed, and each sentence was tokenised using whitespace. The resulting token sequences were compared using SequenceMatcher from Python's difflib module, with autojunk=False, to identify matching and differing segments while preserving their order (Python Software Foundation, n.d.). This approach was suitable for examining translations that shared words but differed in particular expressions, allowing localised substitutions to be identified without requiring identical sentence structures. Candidate correspondences were extracted from segments containing one-to-one substitutions between non-identical word forms, while insertions, deletions and substitutions involving unequal numbers of tokens were excluded. The reproducibility analysis was conducted in a Python 3.13.5 environment, using the standard-library difflib implementation of SequenceMatcher.

To illustrate the procedure, consider the corpus record DLPC-000209, containing the Lutenga sentence "*Abaanaabandhibafangabazaalibwa*" and its Lupakoyo counterpart "*Abaanaabandibafangabazaalibwa*". After preprocessing, the sentences produce the token sequences [*abaana, abandhi, bafa, nga, bazaalibwa*] and [*abaana, abandi, bafa, nga, bazaalibwa*], respectively. The alignment identifies four matching tokens and a single replacement, pairing *abandhi* in Lutenga with *abandi* in Lupakoyo. Because this replacement involves one token from each variety, (*abandhi, abandi*) is retained as a candidate word-form correspondence. The restriction to one-to-one substitutions helps avoid ambiguous alignments involving unequal numbers of tokens, although it may exclude relationships expressed through different grammatical constructions.

The extracted pairs were aggregated across the aligned sentences, with repeated occurrences contributing to the frequency of each correspondence. This produced an inventory of 5,644 distinct candidate word-form pairs, including cases where one form was associated with more than one counterpart in the other variety. Although restricting extraction to one-to-one substitutions reduced some alignment ambiguities, it did not guarantee that the paired words were semantically equivalent. Some correspondences could still reflect differences in word order, grammatical structure or translation choices. During contextual screening, selected low-similarity pairs were examined for such possibilities, including potential alignment errors. The remaining inventory was not individually validated and was therefore treated as a collection of candidate word-form associations. To distinguish recurring associations from isolated differences, the subsequent analysis examined how frequently each pair occurred and the contexts in which it appeared.

2.4 Lexical Correspondence Screening

The extracted word-form correspondences were examined to identify recurring pairs that might represent lexical differences between Lutenga and Lupakoyo. To limit the influence of isolated substitutions, only pairs occurring at least five times were retained for further screening. This reduced the initial inventory of 5,644 distinct pairs to 422 recurrent correspondences. Each retained pair was then assessed for character similarity using the SequenceMatcher ratio, which ranges from 0 to 1, with higher values indicating greater resemblance between written forms. Pairs scoring below 0.40 were selected for closer examination because their substantial differences in written form made them potentially more informative for identifying alternative lexical choices. This left 41

candidates for contextual assessment, although differences in spelling alone could not establish whether the paired words expressed equivalent meanings.

The contextual assessment considered the English source sentences alongside their Lutenga and Lupakoyo translations. Up to five occurrences of each candidate pair were examined, selected at approximately even intervals across the available records to capture variation in usage where sufficient examples existed. The examination focused on whether the paired words conveyed comparable meanings in their respective contexts and whether their differences might instead reflect grammatical structure, translation choices or alignment errors. Eight pairs were provisionally identified as plausible lexical alternatives, five involved potentially related lexical or grammatical forms, and two presented difficulties arising from differences in construction or alignment. The remaining 26 pairs could not be classified confidently from the available contexts. This assessment was conducted by the researchers using the available sentence contexts, without independent review by Lusoga language experts or a formal inter-rater agreement procedure. The eight pairs were therefore treated as candidates for further linguistic investigation rather than confirmed lexical distinctions between the two varieties.

2.5 Character-Level Variation Analysis

Character-level analysis was used to examine differences in the written forms of corresponding Lutenga and Lupakoyo words. Each extracted pair was assigned a similarity score using the SequenceMatcher ratio, calculated using Equation (1):

$$S(L,P) = \frac{2M}{|L|+|P|} \quad (1)$$

where L and P represent the Lutenga and Lupakoyo word forms, respectively; $|L|$ and $|P|$denote their character lengths; and M is the number of matching characters identified by the sequence-matching procedure.Scores range from 0 to 1, with higher values indicating greater resemblance between the written forms.The correspondences were grouped into three mutually exclusive categories based on their character-similarity scores: distant-form ($0 \leq S < 0.40$), intermediate-form ($0.40 \leq S < 0.70$) and close-form ($0.70 \leq S \leq 1.00$). These categories were used to describe the overall distribution of written-form differences rather than to establish morphological or phonological relationships.

To examine recurring character transformations, the analysis focused on correspondences occurring at least five times and having similarity scores of 0.70 or higher. Character-level edit operations were extracted from these pairs, with attention restricted to cases involving a single non-equal edit operation. The resulting transformations were grouped into families according to the characters involved and the type of change observed, including insertions, deletions and substitutions. For each family, the analysis recorded both its total number of observations and the number of distinct word pairs in which it occurred. This distinction allowed transformations that appeared frequently in a small number of pairs to be distinguished from those distributed across a wider range of forms.

Five transformation categories were also examined through a targeted audit: initial *o* insertion, *h* deletion, *dh* → *j* substitution, *gh* → *w* substitution and *f* → *s* substitution. Twelve examples were examined from each category to determine whether the extracted pairs exhibited the indicated character change. A retrospective verification against the complete correspondence inventory reproduced all sixty archived examples in descending order of observation frequency. For initial *o* insertion and *h* deletion, the reconstructed selection was restricted to pairs satisfying the corresponding exact character operation, whereas the three substitution categories included pairs containing the specified character sequences without necessarily satisfying the complete transformation. The audit provided a check on the interpretation of these

computationally identified transformations, although the small, frequency-prioritised selection was not intended to provide comprehensive linguistic validation or establish the productivity or general applicability of the patterns beyond the examined examples.

2.6 Correspondence Consistency Analysis

Correspondence consistency was examined to determine how regularly a Lutenga word form was associated with a particular Lupakoyo counterpart and whether the relationship held in the opposite direction. For each extracted pair, its frequency was compared with the total frequency of substitutions involving each form. Lutenga-to-Lupakoyo consistency was calculated as:

$$C_{L\to P}(l,p) = \frac{f(l,p)}{\sum_{\acute{p}} f(l,\acute{p})} \quad (2)$$

Lupakoyo-to-Lutenga consistency was calculated as:

$$C_{P\to L}(l,p) = \frac{f(l,p)}{\sum_{\acute{l}} f(\acute{l},p)} \quad (3)$$

In these equations, l and $p$represent the Lutenga and Lupakoyo word forms, respectively, while $f(l,p)$ denotes the frequency with which the two forms were extracted as a correspondence. The denominator $\sum_{\acute{p}} f(l,\acute{p})$ represents the total frequency of extracted correspondences involving the Lutenga form l across all its Lupakoyo counterparts $\acute{p}$. Similarly, $\sum_{\acute{l}} f(\acute{l},p)$ represents the total frequency of correspondences involving the Lupakoyo form p across all its Lutenga counterparts $\acute{l}$. Both measures range from 0 to 1, with higher values indicating a stronger association between the corresponding forms within the extracted inventory.

The analysis considered consistency in both directions because a strong association from Lutenga to Lupakoyo does not necessarily imply an equally strong association in reverse. Correspondences were examined at thresholds of 80% and 95%, with a pair meeting a given threshold only when its consistency reached that level in both directions. Mutual-top correspondences were also identified by determining whether each form was the other's most frequently observed counterpart. Where several counterparts shared the highest frequency, the implementation selected the first maximum encountered, making the outcome potentially sensitive to the order in which pairs were processed. Because these measures were calculated from the same corpus used for correspondence extraction, they describe the strength of observed associations rather than independently validating dialect-specific usage. The analysis also explored possible relationships among forms with similar written structures, although these groupings were not treated as linguistically verified.

In this study, correspondence regularity was assessed through four complementary indicators: recurrence frequency, mutual-top association, bidirectional consistency and the distribution of character transformations across distinct word pairs. Recurrence indicates how often a particular correspondence was extracted, mutual-top association identifies whether two forms were each other's most frequent counterpart, and bidirectional consistency measures the relative strength of their association in both directions. Transformation distribution describes how widely a character-edit pattern occurred across different word pairs. These indicators describe different aspects of regularity within the corpus; they were not combined into a single systematicity score and do not independently establish productive linguistic rules or dialect-specific usage.

2.7 Robustness and Sensitivity Checks

Robustness and sensitivity checks were conducted to examine whether the findings changed with different analytical thresholds and alignment methods. The thresholds were selected after an initial exploration of the extracted correspondences rather than being fixed in advance. A minimum frequency of five occurrences was used to focus on recurring pairs, while a character-similarity score below 0.40 helped identify pairs with substantial differences in their written forms. To examine the effect of these choices, we compared minimum recurrence frequencies of 3, 5, 10 and 30 occurrences, together with character-similarity cut-offs of 0.30, 0.40 and 0.50, giving twelve threshold combinations.

We also compared the sequence-based alignment procedure with positional alignment, which pairs words according to their positions in the translated sentences. These comparisons allowed us to examine how the thresholds and alignment method influenced the correspondences identified. The results are presented in Section 3.4.

We then examined how nine selected candidate correspondences were distributed across the training, validation and test partitions prepared for the broader machine translation project. This helped determine whether the pairs appeared in different portions of the corpus or were concentrated in particular subsets. Since all three partitions came from the same translated dataset, their comparison did not provide independent linguistic validation. Rather, it offered another way of examining the distribution of the selected correspondences. These checks helped us assess the sensitivity of the findings to our analytical choices, while recognising that consistency across computational procedures does not necessarily establish a linguistic relationship between the paired forms.

3.0 Results

3.1 Recurrent Lexical Correspondences between Lutenga and Lupakoyo

Comparing the aligned Lutenga and Lupakoyo translations produced 5,644 distinct candidate word-form correspondences, 422 of which occurred at least five times. Among these recurring pairs, 41 had character-similarity scores below 0.40, suggesting differences that went beyond minor changes in written form. Examination of their translation contexts identified eight plausible lexical alternatives, five pairs involving potentially related lexical or grammatical forms, and two cases where differences in construction or alignment made interpretation difficult. The remaining 26 pairs could not be classified confidently from the available evidence.

Table 1: Recurrent lexical correspondence candidates and their translation contexts

Word-form pair (observations)	English source and corpus record	Lutenga translation	Lupakoyo translation	Similarity
empiiya–esente (105)	People sold the stones and got money. DLPC-000396	Abantubaatundaamabaalebaafunaempiiya.	Abantubaatundaamabaalebaafunaesente.	0.154
bendha–bataka (33)	People have likes and dislikes. DLPC-002001	Abantubalina bye bendha ne bye batendha.	Abantubalina bye bataka ne bye batataka.	0.333
musengwa–katonda (29)	Priests are servants of God. DLPC-000081	AbaswezibagheerezabaMusengwa.	AbaffaazabaweereryabaKatonda.	0.267
buti–atyano (28)	Women are now represented in parliament. DLPC-000193	Abakyalabutibakiikirirwa mu lukiikolw'eighangaolukulu.	Abakyalaatyanobakiikirirwaomulukiikolw'eih angaolukulu.	0.200
ghazira–wabula (22)	People are being arrested without justifiable reasons. DLPC-001309	Abantubagemebwangaghaziransongadhi matiza.	Abantubagemebwangawabulansongagimatizya.	0.308
ghalala–hamo (18)	We pray together every day. DLPC-004003	Tusabiraghalalabulilunaku.	Tusabirahamobulilunaku.	0.364
mpiiya–sente (17)	Most fundraisings are money-oriented. DLPC-000224	Okusondaokusingakuluubirirampiiya.	Okusondaokusingakuluubirirasente.	0.000
leero–lwati (16)	They called for a meeting today. DLPC-000996	Beeseolukiikoleero.	Beetereolukiikolwati.	0.200

Note: Frequencies represent extracted aligned observations. The pairs are provisional lexical candidates and have not been independently validated. Their inclusion does not imply exclusive use in either variety.

The examples in Table 1 suggest several possible vocabulary differences between Lutenga and Lupakoyo. The pairs *empiiya–esente* and *mpiiya–sente* both refer to money and may represent related forms of the same lexical correspondence. Other pairs suggest alternative expressions for liking or preference (*bendha–bataka*), God (*musengwa–katonda*), the present time (*buti–atyano*), absence (*ghazira–wabula*), togetherness (*ghalala–hamo*) and today (*leero–lwati*). These interpretations are based on the meanings conveyed in the accompanying translations and should therefore be treated with caution. Some differences may arise from grammatical constructions or individual translation choices rather than established vocabulary distinctions between the two varieties. The eight pairs consequently provide candidates for further linguistic investigation, not a definitive inventory of dialect-specific lexical differences.

3.2 Recurrent Form-Level Variation

The 5,644 extracted pairs also varied in how closely their written forms resembled one another. Most were classified as close-form correspondences, accounting for 4,266 pairs (75.6%), compared with 1,116 (19.8%) in the intermediate-form category and 262 (4.6%) in the distant-form category. Thus, although the alignment identified differences between the two varieties, many of the paired words retained considerable character similarity. A closer examination of these differences identified 52 character-edit families, several of which recurred across multiple word pairs. Table 2 presents the most frequent transformations.

Table 2: Selected recurrent character-edit families

Lutenga → Lupakoyo edit	Observations	Distinct word pairs
Insertion of o	2,699	10
Deletion of h	554	12
Deletion of g	526	31
Insertion of y	375	30
Insertion of e	284	7
Insertion of b	255	10
dh → j	220	14
gh → w	185	3

Note: Observations represent extracted character-edit occurrences rather than independent linguistic rules. Repeated word forms may contribute multiple observations. The direction of each edit reflects the computational comparison from Lutenga to Lupakoyo and does not imply a historical direction of linguistic change

The frequency of a transformation did not always correspond to the number of different word pairs in which it appeared. Insertion of o, for example, accounted for 2,699 observations but occurred in only ten distinct pairs. Deletion of g was much less frequent, with 526 observations, yet appeared across 31 pairs. A similar difference was observed between insertion of y, which occurred 375 times across 30 pairs, and insertion of e, which contributed 284 observations from just seven pairs. Some transformations were therefore frequent because particular word pairs appeared repeatedly, while others were distributed across a wider range of forms.

The targeted audit provided a closer look at how well selected transformations accounted for these written differences. All twelve examples examined for word-initial *o* insertion and all twelve for *h* deletion could be explained by their respective string-transformation rules. The results were less consistent for the other transformations: nine of the twelve examples involving *dh* → *j* satisfied the specified rule, compared with four for *gh* → *w* and three for *f* → *s*. In these latter cases, several pairs contained additional character differences that could not be explained by the transformation alone. While some recurring form differences could therefore be described through a single character

operation, others were more complex. Whether these written patterns reflect broader phonological or morphological relationships remains a question for further linguistic investigation.

3.3 Systematicity and Recurrence of Cross-Variety Correspondences

Of the 5,644 candidate correspondences, 422 appeared at least five times, 193 occurred at least ten times, and 53 were recorded thirty or more times. The 422 recurring pairs included 277 close-form, 104 intermediate-form and 41 distant-form correspondences, indicating that repeated associations were found among both closely resembling words and forms with substantial written differences.

The next question was how consistently individual words were associated with their counterparts across the translations. Among the 422 recurring pairs, 369 (87.4%) were mutual top correspondences,meaning that each form was identified as a most frequently extracted substitution partner of the other, under the tie-handling procedure described in Section 2.6. Strong associations were also evident from the bidirectional consistency measures: 266 pairs (63.0%) reached at least 80% consistency in both directions, while 184 (43.6%) reached 95% or higher. These results are summarised in Table 3.

Table 3: Recurrence and consistency of extracted correspondences

Indicator	Candidate pairs
Recurring at least five times	422
Recurring at least ten times	193
Recurring at least thirty times	53
Mutual top correspondences	369
At least 80% bidirectional consistency	266
At least 95% bidirectional consistency	184

Note: Consistency is calculated from extracted substitution partners and should not be interpreted as the probability that speakers of either variety will use a particular form

We also examined whether the treatment of ties between equally frequent counterparts influenced the mutual-top results. The original procedure identified 369 mutual-top pairs among the 422 recurring correspondences. Across the full inventory, 125 Lutenga forms and 214 Lupakoyo forms had equally frequent top counterparts, but only 13 recurring pairs involved such ties. Allowing all equally frequent counterparts to count as top matches increased the number of mutual-top pairs to 373, while requiring a unique top counterpart in both directions reduced it to 363. These differences were relatively small, suggesting that the overall pattern of reciprocal associations remained largely unchanged across the three approaches. Nevertheless, the exact number of mutual-top pairs depended on how ties were handled. Not all frequently occurring pairs showed the same degree of consistency. Some forms were repeatedly paired with one particular counterpart, while others appeared alongside several alternatives. The character-edit analysis also showed that recurring transformations were distributed across different numbers of word pairs, as reported in Table 2. These counts describe the distribution of written-form differences in the extracted correspondences, but do not establish that the pairs belong to independent lexical families or that the transformations represent productive linguistic rules.

The results show that several Lutenga–Lupakoyo correspondences recur across the translations, often with a strong association between particular word forms. However, these associations vary in consistency and distribution: some are concentrated in a few frequently occurring words, while others extend across a wider range of forms. Their recurrence provides evidence of patterned variation in the

analysed corpus, although further linguistic investigation would be needed to determine how far these patterns extend beyond the translations examined here.

3.4 Robustness and Sensitivity of the Extracted Correspondences

The number of distant-form correspondence candidates varied depending on the frequency and character-similarity thresholds applied. Table 4 shows the results for the twelve combinations examined.

Under the main analytical setting of at least five observations and a similarity score below 0.40, 41 candidates were identified. Keeping the frequency threshold unchanged but lowering the similarity cut-off to 0.30 reduced this number to 14, while raising it to 0.50 increased it to 67. The number also varied with the frequency threshold, falling to 18 when the minimum was increased to ten observations and rising to 61 when it was lowered to three, with the similarity cut-off held at 0.40.

Table 4: Number of distant-form correspondence candidates under different frequency and similarity thresholds

Minimum recurrence frequency	Similarity < 0.30	Similarity < 0.40	Similarity < 0.50
3	22	61	105
5	14	41	67
10	8	18	31
30	2	6	9

Note. Values represent the number of extracted correspondence pairs meeting both the minimum recurrence frequency and the specified character-similarity cut-off.

The alignment method also influenced the results, with sequence-based alignment producing 5,644 distinct correspondence pairs compared with 12,301 from positional alignment. Of the 50 most frequent correspondences identified by each alignment method, 39 pairs were common to both lists, indicating some agreement in frequently occurring associations despite the considerable difference between the overall inventories. We also examined nine selected correspondences across the training, validation and test partitions to establish how they were distributed within the corpus. Although this provided another view of the selected pairs, the partitions came from the same collection of translations and could not offer independent linguistic validation.

The sensitivity checks showed that the number and composition of extracted correspondences depended partly on the thresholds and alignment procedure used. While the recurring associations provide useful evidence of variation within the corpus, their interpretation remains tied to these analytical choices. They therefore offer a basis for examining lexical and written-form differences between Lutenga and Lupakoyo rather than a complete or independently validated account of those differences.

4.0 Discussion

4.1 Lexical Differentiation beyond Minor Written Variation

Some of the recurring differences between Lutenga and Lupakoyo involved words with substantially different written forms. Pairs such as *empiiya–esente*, *bendha–bataka* and *musengwa–katonda* appeared repeatedly despite sharing relatively few characters, suggesting that the relationship between the two varieties cannot be understood through written-form similarity alone. Although aggregate similarity measures can indicate how closely related language varieties are, they may conceal differences in the vocabulary used to express comparable meanings.

These findings are consistent with the emphasis placed on fine-grained linguistic variation in Bantu languages by Bloom Ström and Petzell (2025). However, interpreting individual correspondences requires attention to the relationships between the forms themselves. For example, *empiiya–esente* and *mpiiya–sente* may represent related forms of the same underlying

correspondence rather than separate lexical distinctions. Other pairs may involve grammatical differences that are difficult to identify from isolated words, particularly in Bantu languages, where morphological structure can complicate the interpretation of surface forms (Edelsten et al., 2024). Kuparinen and Scherrer (2024) similarly found that the patterns identified in dialect data depend partly on whether the analysis considers whole words, character sequences or segmented forms.

The study also provides a different perspective on corpus-based research involving Lusoga. Whereas de Schryver and Nabirye (2022) examined lexical development through neology detection, the present analysis focuses on differences between corresponding expressions in Lutenga and Lupakoyo. Examining recurrent correspond-dences within their translation contexts helps identify plausible vocabulary alternatives while also revealing cases in which grammatical constructions or individual translation choices complicate interpretation. This is particularly useful where paired words have substantially different written forms, since character similarity alone cannot establish whether they express comparable meanings. The resulting inventory therefore provides a more focused starting point for investigating lexical differentiation than the complete set of computationally extracted pairs, while leaving the linguistic status of individual correspondences open to further investigation.

4.2 Recurrent Written-Form Correspondences

The character-edit analysis revealed an important distinction between how often a written-form difference occurred and how widely it appeared across different words. Some transformations occurred frequently because particular word pairs appeared repeatedly, whereas others were distributed across a wider range of forms. High observation frequency, therefore, did not necessarily indicate that a transformation was widespread across the vocabulary. Looking at both its frequency and the number of different word pairs in which it appears therefore helps explain how widely a particular written-form difference is represented in the corpus.

The targeted rule-specificity audit showed that some written-form differences were easier to account for than others. While certain word pairs differed through a single character operation, others contained additional changes that could not be explained by the identified transformation alone. This makes it difficult to treat recurring character-edit patterns as general linguistic rules without examining the words more closely. As Edelsten et al. (2024) illustrate in their work on Bantu morphosyntactic variation, understanding relationships between forms requires attention to their linguistic structure, not simply their written resemblance. Similar concerns arise in research on Icibemba orthography, where variation in written representation presents challenges for computational language processing (Kasonde, 2025).

These differences may also matter when developing computational resources for closely related varieties. Lin et al. (2024), for example, reported improvements in sentiment analysis and machine translation when orthographic alternatives were incorporated into Nigerian Pidgin data. For Lutenga and Lupakoyo, the observed transformations provide candidate written-form correspondences that may reflect differences in spelling conventions, morphological structure or individual translation choices. The present analysis cannot reliably distinguish among these explanations, but the distribution of transformations across distinct word pairs helps identify which patterns warrant closer linguistic examination. Their potential relevance to computational language processing therefore remains a question for subsequent empirical investigation rather than a demonstrated benefit of this study.

4.3 Recurrence and Consistency of Cross-Variety Correspondences

The findings show that recurring word-form correspondences between Lutenga and

Lupakoyo differ in how consistently the paired forms occur together. Some words were repeatedly associated with one particular counterpart, while others appeared with several alternatives. This distinction is important because a frequently occurring pair does not necessarily represent a strong or stable association. Examining recurrence alongside bidirectional consistency therefore helps distinguish pairs that occur often from those whose associations remain relatively consistent within the corpus.

These differences also show why correspond-dence regularity cannot be judged from a single measure. As discussed by Bloom Ström and Petzell (2025), examining closely related Bantu varieties at a finer level can reveal variation that broader comparisons may overlook. In the present study, recurrence, bidirectional consistency and the distribution of character transformations each captured a different aspect of the observed relationships. In particular, character transformations that recur frequently may still be concentrated in only a few distinct word forms. Considering these measures together provides a more detailed account of the patterns represented in the corpus, without assuming that they reflect productive morphological processes. Establishing whether such processes are involved would require closer examination of the forms and their grammatical relationships, as illustrated byEdelsten et al. (2024).

The shared English source sentences made it possible to compare Lutenga and Lupakoyo translations within similar contexts, but they also introduced an important consideration. The correspondences reflect choices made by translators working with a particular collection of sentences, and some associations may have been influenced by the vocabulary and constructions represented in those sources. Shadrova (2025) similarly demonstrates that lexical choices can vary across productions under comparable task conditions, providing a useful reminder that recurring word associations need not reflect fixed linguistic preferences. The observed regularity is therefore meaningful within the corpus, but its wider distribution among speakers and across communicative settings remains to be established.

The sensitivity checks also highlighted the influence of analytical choices on the correspondences identified. Changing the recurrence and similarity thresholds altered the number of distant-form candidates, while the two alignment methods produced substantially different inventories despite sharing several frequently occurring pairs. This suggests that the extracted correspondences should be understood in relation to the procedures used to identify them. Their recurrence within the corpus remains informative, but the precise inventory is not independent of the analytical method.

4.4 Implications for Computational Processing of Closely Related Varieties

The recurring lexical and written-form differences identified in this study have implications for how closely related varieties are represented in computational resources. Although Lutenga and Lupakoyo share considerable linguistic structure, treating their translations as interchangeable may overlook distinctions that recur within comparable contexts. Retaining separate variety labels in parallel datasets would therefore allow researchers to investigate recurring vocabulary and written-form differences while preserving the substantial similarities shared by Lutenga and Lupakoyo. This concern is consistent with research on machine translation into low-resource language varieties, which has examined the challenges of representing and generating regional varieties rather than treating them as interchangeable with more widely represented forms (Kumar et al., 2021). At the same time, the predominance of close-form correspondences suggests that the two varieties share substantial written-form similarities, making it equally

important to account for what they have in common.

Previous research offers different perspectives on the computational importance of such variation. Lin et al. (2024) demonstrated the potential value of representing orthographic alternatives in Nigerian Pidgin data, while Hopton and Aepli (2024) found that a multilingual model trained on several Occitan dialects remained relatively robust to dialectal variation, suggesting that spelling normalisation may not always be necessary.These findings suggest that the effect of written-form variation depends on the language, the computational task and how the training data are represented. Faisal et al. (2024) also reported differences in NLP performance across dialects and closely related languages, suggesting that results obtained for one variety should not automatically be assumed to hold for another. The present study did not test these effects for Lutenga and Lupakoyo, so the potential benefits of explicitly modelling the identified correspondences remain an empirical question.

For future Lusoga language-resource development, the findings point to several areas worth investigating. The extracted correspondences could help researchers identify vocabulary that deserves closer examination in each variety, while the recurring character differences offer examples for investigating possible spelling and morphological variation. Keeping the two varieties separately labelled would also make it possible to examine whether the same correspondences appear in other collections of Lusoga texts and whether they remain consistent across different speakers and contexts. Such resources could eventually support the development and evaluation of variety-aware machine translation models, although the present study does not establish whether incorporating these patterns would improve translation performance. This focus on relationships between regional varieties also extends the scope of corpus-based Lusoga research beyond the lexical development examined by de Schryver and Nabirye (2022).

4.5 Limitations of the Study

The study drew on 14,597 aligned English–Lutenga–Lupakoyo sentence triplets produced through translation rather than naturally occurring conversations or texts. Using common English source sentences provided comparable contexts for examining the two varieties, but the resulting translations may also reflect individual vocabulary preferences, the wording of the source sentences and the types of expressions represented in the corpus. Wang and Xin (2024) similarly identify textual characteristics and translator-related factors as important considerations when interpreting corpus-based translation findings. The findings consequently describe variation within this collection and may not capture the full range of Lutenga and Lupakoyo usage across different speakers, communities and communicative situations.

There are also limitations associated with the computational procedures. Sequence alignment identifies corresponding word positions and substitutions within translated sentences, but the resulting pairs are not necessarily direct linguistic equivalents, particularly when the translations differ in word order or grammatical construction. Character-similarity scores and edit operations likewise describe relationships between written forms without establishing whether the differences arise from vocabulary choice, spelling practices, morphology or phonology. The targeted rule-specificity audit helped examine selected character transformations, but it could not establish their wider linguistic significance.

The eight pairs identified as possible lexical alternatives also require further linguistic examination. Their selection was based on the English source sentences and the corresponding Lutenga and Lupakoyo translations, which provided useful context for comparing their meanings. However, the assessment was carried out by the researchers without independent

validation from native-speaker experts. Although the paired expressions appeared to convey comparable meanings in the contexts examined, it remains unclear whether they reflect established differences between the two varieties or simply the choices made by individual translators. The screening procedure may also have overlooked other lexical differences, particularly those that did not meet the frequency and character-similarity thresholds. The eight pairs should therefore be regarded as candidates for further investigation rather than confirmed lexical distinctions between Lutenga and Lupakoyo.

The recurrence and consistency measures were derived from the same corpus and may therefore be influenced by repeated expressions or relatively stable choices made during translation. Although some character transformations occurred across several distinct word pairs, this distribution alone does not establish that the patterns represent productive morphological processes or general linguistic rules. Further examination of naturally occurring language data, supported by targeted native-speaker judgements and more detailed linguistic annotation, would help establish whether the observed correspondences extend beyond the translated corpus and clarify their linguistic significance.

5.0 Conclusion

This study examined lexical and written-form variation between Lutenga and Lupakoyo through a computational comparison of aligned translations. The analysis identified recurring word-form correspondences that differed in their frequency, written-form similarity and consistency across the corpus. Some correspondences involved substantially different words, while others reflected recurring character-level differences whose frequency did not always correspond to their distribution across distinct word pairs. Examining these relationships alongside bidirectional consistency provided a more detailed account of the variation represented in the corpus than overall similarity measures alone.

The study contributes an inventory of computationally extracted candidate correspondences and an account of how these associations recur and vary between the two Lusoga varieties. Although several correspondences occurred repeatedly and showed relatively strong associations, the findings do not establish that particular words are exclusive to either variety or that the observed character changes represent general linguistic rules. Further investigation using naturally occurring language data, native-speaker judgements and closer linguistic examination would help clarify the status and wider distribution of these patterns. The extracted correspondences may inform future investigations of variety-sensitive Lusoga language resources, although their usefulness for machine translation and other computational tasks remains to be established empirically.

Declarations

Funding

This research received no external funding.

Ethics Approval

The study forms part of an ongoing doctoral research project approved by the relevant institutional and national research authorities in Uganda, including the Higher Degrees and Research Ethics Committee (HAUREC) and the Uganda National Council for Science and Technology (UNCST). Permission to conduct research in the Busoga sub-region was also obtained from the Busoga Kingdom. The present analysis used an existing English–Lutenga–Lupakoyo parallel corpus developed under the approved research project and did not involve additional participant recruitment or primary data collection.

Consent to Participate

Informed consent was obtained from participants involved in the original corpus

development, in accordance with the approved research protocol.

Consent for Publication

Not applicable. The manuscript contains no identifiable personal information relating to individual research participants.

Data Availability

The English–Lutenga–Lupakoyo parallel corpus and the derived analytical datasets are not publicly available at present because they form part of an ongoing doctoral research project. Requests for access may be directed to the corresponding author and will be considered in accordance with applicable ethical, institutional and data-sharing requirements.

Code Availability

The Python code used for the computational analysis, together with the supporting analytical audit files and reproducibility instructions, is available from the corresponding author upon reasonable request. Access to the underlying corpus is subject to applicable institutional and data-sharing requirements.

Conflicts of Interest

The authors declare no conflicts of interest.

References

Bloom Ström, E.-M., & Petzell, M. (2025). Micro-variation approaches to Bantu language varieties. In L. Marten, E. Hurst-Harosh, N. C. Kula, & J. Zeller (Eds.), *The Oxford guide to the Bantu languages* (pp. 423–432). Oxford University Press. https://doi.org/10.1093/oso/9780198808343.003.0033

Creissels, D. (2025). Typology and variation. In L. Marten, E. Hurst-Harosh, N. C. Kula, & J. Zeller (Eds.), *The Oxford Guide to the Bantu Languages* (pp. 28–42). Oxford University Press. https://doi.org/10.1093-/oso/9780198808343.003.0004

de Schryver, G.-M., & Nabirye, M. (2018). Corpus-driven Bantu lexicography part 1: Organic corpus building for Lusoga. *Lexikos*, *28*, 32–78. https://doi.org/-10.5788/28-1-1457

de Schryver, G.-M., & Nabirye, M. (2022). Towards a monitor corpus for a Bantu language: A case study of neology detection in Lusoga. In A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs, & P. Storjohann (Eds.), *Dictionaries and society: Proceedings of the 20th EURALEX International Congress* (pp. 833–849). IDS-Verlag. https://doi.org/10.14618/-phpy-6r66

Edelsten, P., Guérois, R., & Marten, L. (2024). Comparative analysis of morphosyntactic variation in Bantu languages: Parameters, data representation, and design. In E.-M. Bloom Ström, H. Gibson, R. Guérois, & L. Marten (Eds.), *Morphosyntactic variation in Bantu* (pp. 25–52). Oxford University Press. https://doi.org/10.1093/oso/-9780198821359.003.0002

Faisal, F., Ahia, O., Srivastava, A., Ahuja, K., Chiang, D., Tsvetkov, Y., & Anastasopoulos, A. (2024). DialectBench: A NLP benchmark for dialects, varieties, and closely related languages. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14412–14454. https://doi.org/10.18653/v1/2024.acl-long.777

Hopton, Z. W., & Aepli, N. (2024). Modeling orthographic variation in Occitan's dialects. *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, 78–88. https://doi.org/10.18653/v1/-2024.vardial-1.6

Isiko, A. P. (2018). *Gender roles in traditional healing practices in Busoga* [Doctoral dissertation, Leiden University]. http://hdl.handle.net/1887/63215

Joshi, A., Dabre, R., Kanojia, D., Li, Z., Zhan, H., Haffari, G., & Dippold, D. (2025). Natural language processing for dialects of a language: A survey. *ACM Computing Surveys*, *57*(6), Article 149, 1–37. https://doi.org/10.1145/3712060

Kasonde, A. (2025). The long march to Unicode: A digital approach to variability in Icibemba orthography. *Cogent Arts & Humanities*, *12*(1), 2477347. https://doi.org/10.1080/23311983.2025.2477347

Kimera, R., Rim, D. N., & Choi, H. (2022). Building a parallel corpus and training translation models between Luganda and English. *Journal of KIISE*, *49*(11), 1009–1016. https://doi.org/10.5626/jok.2022.49.11.1009

Kumar, S., Anastasopoulos, A., Wintner, S., & Tsvetkov, Y. (2021). Machine translation into low-resource language varieties. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 110–121. https://doi.org/10.18653/v1/2021.acl-short.16

Kuparinen, O., & Scherrer, Y. (2024). Corpus-based dialectometry with topic models. *Journal of Linguistic Geography*, *12*(1), 1–12. https://doi.org/10.1017/jlg.2024.6

Lin, P., Scholman, M., Saeed, M., & Demberg, V. (2024). Modeling orthographic variation improves NLP performance for Nigerian Pidgin. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 11510–11522. https://aclanthology.org/2024.lrec-main.1006/

Min, J., Tan, T., & Malancon, B. (2024). Hybrid distance-statistical-based phrase alignment for analyzing parallel texts in standard Malay and Malay dialects. *Malaysian Journal of Computer Science*, *37*(1), 1–25. https://doi.org/10.22452/mjcs.vol37no1.1

Nabirye, M. (2016). *A corpus-based grammar of Lusoga* [Doctoral dissertation, Ghent University]. http://hdl.handle.net/1854/-LU-8125501

Nabirye, M., de Schryver, G.-M., & Verhoeven, J. (2016). Lusoga (Lutenga). *Journal of the International Phonetic Association*, *46*(2), 219–228. https://doi.org/10.1017/S0025-100315000249

Python Software Foundation. (n.d.). *difflib — Helpers for computing deltas*. Retrieved November 7, 2026, from https://-docs.python.org/3.13/library/difflib.html

Shadrova, A. (2025). No three productions alike: Lexical variability, situated dynamics, and path dependence in task-based corpora. *Open Linguistics*, *11*(1), Article 20240036. https://doi.org/10.1515/opli-2024-0036

Shim, R. S., & Plank, B. (2025). Dialetto, ma quanto dialetto? Transcribing and evaluating dialects on a continuum. *Findings of the Association for Computational Linguistics: NAACL 2025*, 838–849. https://doi.org/10.18653/v1/-2025.findings-naacl.48

Wang, G., & Xin, Y. (2024). An analytical framework for corpus-based translation studies. *Humanities and Social Sciences Communications*, *11*(1), 1709.

https://doi.org/10.1057/s41599-024-04250-4

Xie, R., Ahia, O., Tsvetkov, Y., & Anastasopoulos, A. (2024). Extracting lexical features from dialects via interpretable dialect classifiers. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 54–69. https://doi.org/10.18653/-v1/2024.naacl-short.5