

Publication status: This preprint has been published elsewhere.

DOI of the published preprint: <https://doi.org/10.5281/zenodo.23277947>

Beyond Surface Similarities: A Corpus-Based Comparison of Tamil and Korean

Jinwoo Chae

<https://doi.org/10.1590/SciELOPreprints.18437>

Submitted on: 2026-10-16

Posted on: 2026-10-23 (version 1)

(YYYY-MM-DD)

The moderation of this preprint received the endorsement of:

- Jihye Chae (ORCID: <https://orcid.org/0009-0006-3432-5112>)

Beyond Surface Similarities: A Corpus-Based Comparison of Tamil and Korean

[Jinwoo Chae ] <https://orcid.org/0009-0000-4586-605X>

[Doctor of Business Administration], GlobalNxt University, Malaysia, ROR: 00kb5s826

Abstract

Tamil and Korean share several typological characteristics, including predominantly subject–object–verb (SOV) word order, agglutinative morphology, and the use of grammatical markers to express relationships between sentence constituents. Nevertheless, similarities in linguistic structure do not necessarily indicate common historical ancestry. This study examines lexical, morphological, syntactic, and semantic relationships between Tamil and Korean through a comparative corpus-based framework. The analysis focuses on lexical correspondence, constituent order, case marking, verbal morphology, negation, and honorific expressions. Levenshtein similarity, morphological annotation, syntactic coding, and comparisons with control languages provide complementary approaches to evaluating linguistic similarities. The comparative analysis demonstrates the importance of distinguishing shared grammatical functions from identical linguistic forms. Although both languages employ suffixation and frequently place verbs at the end of clauses, their lexical systems and grammatical constructions exhibit important differences. The study further establishes that lexical resemblance must be evaluated through semantic correspondence, regular phonological patterns, and historical evidence rather than superficial similarity alone. By integrating quantitative linguistic measures with qualitative grammatical interpretation, this research provides a systematic framework for investigating linguistic similarities while avoiding unsupported assumptions about language origins. The findings emphasize that typological comparison can reveal meaningful structural parallels without independently establishing a genealogical relationship between Tamil and Korean.

Keywords: Tamil language; Korean language; corpus linguistics; comparative linguistics; agglutinative morphology; syntactic typology; lexical similarity; historical linguistics

Chapter 1. Introduction

1.1. Background of the Study

Comparative linguistics examines similarities and differences among languages to understand their grammatical organization, lexical development, and historical relationships. Languages may share particular structural characteristics despite differences in their vocabularies, writing systems, and historical development. These similarities may result from general linguistic tendencies, independent development, language contact, or common ancestry. Distinguishing among these explanations requires systematic examination of linguistic evidence.

Tamil belongs to the Dravidian language family and has a long literary tradition. Its grammatical system includes extensive suffixation, grammatical case marking, and predominantly verb-final sentence structures. Korean similarly employs extensive morphological marking, grammatical particles, and predominantly subject–object–verb ordering in ordinary declarative clauses. Both languages can express grammatical relationships through morphological elements attached to lexical bases, although the distribution and functions of these elements differ.

The following comparison demonstrates how similar meanings can be expressed through different linguistic forms.

Table 1. Comparative Linguistic Features of Tamil and Korean Through Sentence-Level Analysis

Linguistic feature	Tamil	Korean
Original sentence	நான் சோறு சாப்பிட்டேன்.	나는 밥을 먹었다.
Transliteration	Nāṇ cōru cāppittēṇ.	Naneun babeul meogeotda.
English translation	I ate rice.	I ate rice.
Basic constituent order	Subject–Object–Verb	Subject–Object–Verb
Subject expression	நான் (nāṇ), “I”	나 (na), “I”; 나는 (naneun) with topic marking
Object expression	சோறு (cōru), “rice/food”	밥을 (babeul), “rice/meal” with object marking
Verb position	Final	Final

The two sentences exhibit a comparable constituent order, but their lexical items and

grammatical marking are not identical. Korean *neun* marks the topic in this sentence, whereas the Tamil subject pronoun does not carry an equivalent overt topic suffix in the construction shown. Korean *eul* marks the object, while the Tamil object noun in this sentence lacks the corresponding overt accusative form (Shanmugamoorthy et al., 2022).

These differences illustrate why linguistic comparison requires more than identifying broad structural similarities. Researchers must examine how grammatical functions are expressed, how frequently particular constructions occur, and whether similarities remain consistent across different linguistic contexts.

A corpus-based approach provides a systematic means of investigating these questions. By examining collections of naturally occurring language data, researchers can identify recurring patterns, compare their distributions, and distinguish general structural characteristics from similarities limited to particular expressions (Abdullah, 2009).

1.2. Research Problem

Comparisons between languages frequently emphasize selected lexical resemblances or shared grammatical characteristics. However, the interpretation of these similarities presents several methodological challenges.

First, lexical resemblance does not necessarily indicate historical relatedness. Short words may resemble one another by chance, and differences in transcription can influence the apparent similarity of forms (Yeo et al., 2022a).

Second, typological characteristics such as SOV word order and agglutinative morphology occur across numerous languages. Their presence in two languages does not independently establish a common historical origin.

Third, linguistic comparisons can be affected by corpus composition, translation practices, morphological segmentation, and the selection of lexical items. A comparison based on inconsistent data may exaggerate similarities or conceal important differences (Thanabalasingam, 2023).

The central research problem concerns how lexical and grammatical similarities between Tamil and Korean can be examined systematically while distinguishing observable linguistic

patterns from claims about their historical origins.

1.3. Research Objectives and Questions

The study has four objectives:

1. To examine lexical similarities and differences between Tamil and Korean using systematic comparison criteria.
2. To analyze the morphological and syntactic characteristics of the two languages, particularly word order, case marking, verbal morphology, and grammatical particles.
3. To establish a comparative framework for evaluating linguistic similarities against appropriate control languages.
4. To distinguish surface resemblance, structural similarity, and evidence relevant to historical linguistic relationships.

The research addresses four questions:

- RQ1: What lexical similarities and differences can be identified between Tamil and Korean?
- RQ2: How do the two languages compare in morphological organization, constituent order, and grammatical marking?
- RQ3: To what extent are their shared structural characteristics distinctive when compared with other languages?
- RQ4: What evidence is necessary to distinguish typological similarities from possible historical relationships?

1.4. Significance of the Study

This study contributes to comparative linguistics by integrating lexical analysis with grammatical and syntactic comparison. Its analytical framework emphasizes predefined comparison criteria, phonological normalization, grammatical annotation, and the evaluation of alternative explanations (Arokiyaraj, 2020).

The study also demonstrates the importance of distinguishing linguistic description from historical interpretation. Similarities can be documented without assuming that they arise from

a common ancestor, and differences can be analyzed without disregarding meaningful structural parallels.

This approach provides a basis for more rigorous investigations of language pairs whose similarities have attracted linguistic and public interest.

Chapter 2. Literature Review and Conceptual Framework

2.1. Corpus Linguistics and Comparative Language Research

Corpus linguistics examines language through systematically collected textual or spoken data. Depending on the research objective, corpus analysis can investigate lexical frequencies, collocations, morphological patterns, grammatical constructions, and syntactic distributions.

Two corpus types are particularly relevant to comparative research. A comparable corpus contains texts from different languages selected according to similar genres, periods, or communicative purposes. A parallel corpus contains source texts and their translations into another language (Vasuthavan & Sivanadhan, 2025).

Comparable corpora are useful for examining language-specific patterns without relying exclusively on translated material. Parallel corpora facilitate the examination of how equivalent meanings are expressed through different grammatical systems.

For Tamil and Korean, combining these approaches makes it possible to investigate both naturally occurring linguistic patterns and the grammatical choices used to express comparable meanings. Such comparisons must account for differences in genre, register, text length, and morphological segmentation (Arokiyaraj, 2026).

2.2. Lexical Similarity and Historical Relatedness

Lexical comparison examines the relationship between word forms and their meanings. Relevant dimensions include phonological resemblance, semantic correspondence, morphological structure, and historical documentation.

Computational measures such as Levenshtein distance quantify the difference between two strings. However, the resulting scores measure formal similarity rather than historical relatedness. A high score cannot independently establish that two words share a common origin.

Historical comparison requires additional evidence, particularly regular sound correspondences across multiple lexical items, plausible semantic development, and historical forms consistent with the proposed relationship. Borrowing and accidental resemblance must also be considered.

A systematic comparison should therefore use a predefined vocabulary list rather than selecting words after their resemblance has been noticed. This procedure reduces selection bias and makes the analysis more reproducible.

2.3. Agglutinative Morphology and Syntactic Typology

Agglutinative languages frequently combine lexical stems with grammatical morphemes to express meanings such as case, tense, person, number, and other grammatical distinctions. Tamil and Korean both display substantial morphological marking, but their grammatical systems differ in important respects (Yeo et al., 2022b).

Table 2. A Comparative Analysis of Grammatical Features in Tamil and Korean

Grammatical feature	Tamil	Korean	Comparative interpretation
Basic constituent order	Predominantly SOV	Predominantly SOV	Shared typological tendency
Case marking	Case suffixes on nouns and noun phrases	Case particles attached to noun phrases	Comparable grammatical functions with different forms and distributions
Topic marking	Topic and discourse functions expressed through context and available constructions	Overt topic particles, including 은/는 (eun/neun)	Differences in grammatical realization
Verbal morphology	Suffixation expresses tense, agreement, mood, and other distinctions	Verb endings express tense, mood, speech level, and other distinctions	Morphological functions require detailed comparison
Negation	Different negative constructions according to grammatical context	Different negative constructions according to grammatical context	Construction-level comparison is necessary
Honorifics and	Socially conditioned	Extensive honorific	Similar

politeness	honorific and politeness expressions	and speech-level distinctions	communicative functions do not imply identical systems
------------	--------------------------------------	-------------------------------	--

The comparison indicates that the two languages share certain broad typological properties while differing in the specific mechanisms through which grammatical information is encoded. These distinctions are essential for avoiding the assumption that languages with similar word order or morphological complexity necessarily share a historical origin.

2.4. Conceptual Framework

The study distinguishes three dimensions of linguistic similarity.

Surface Similarity

Lexical forms that resemble one another in pronunciation or written representation.

Structural Similarity

Recurring patterns in constituent order, morphological organization, grammatical functions, and sentence construction.

Historical Relatedness

Relationships supported by regular sound correspondences, historically grounded lexical evidence, and other independent evidence of linguistic descent or contact.

These dimensions are related but analytically distinct. Surface similarity can motivate further investigation, and structural similarity can reveal typological parallels. Historical relatedness, however, requires evidence that cannot be inferred from either form of resemblance alone (Bharath & Madhvanath, 2007).

Chapter 3. Research Methodology

3.1. Research Design

The study employs a corpus-based comparative design integrating quantitative and qualitative linguistic analysis. Quantitative methods are appropriate for measuring lexical resemblance and comparing the distribution of grammatical constructions. Qualitative analysis is necessary for interpreting morphological functions, semantic relationships, and contextual

differences.

The research design incorporates four analytical components:

1. Corpus collection and linguistic data normalization.
2. Lexical comparison using phonological and computational measures.
3. Morphological and syntactic annotation using a predefined codebook.
4. Evaluation of observed similarities against control languages and alternative explanations.

The analytical procedure prioritizes consistent inclusion criteria and transparent documentation. These requirements are necessary to ensure that the results can be reproduced and evaluated independently.

3.2. Corpus Selection and Data Collection

The comparison requires Tamil and Korean texts selected according to comparable genre, period, register, and communicative purpose. Contemporary informational prose and news texts provide useful material for examining common vocabulary and grammatical constructions. Literary texts may be examined separately because stylistic conventions can influence lexical and syntactic distributions (Cho & Park, 2024).

The following corpus structure provides a consistent basis for organizing the comparison.

Table 3. Corpus Design and Data Components for the Comparative Analysis of Tamil and Korean

Corpus component	Tamil data	Korean data	Analytical purpose
Contemporary informational texts	Matched textual collection	Matched textual collection	General vocabulary and grammatical patterns
News and public-affairs texts	Matched textual collection	Matched textual collection	Formal register and recurring constructions
Parallel sentences	Source or translated sentences	Corresponding aligned sentences	Comparison of equivalent meanings
Basic vocabulary	Standardized lexical entries	Corresponding lexical entries	Controlled lexical comparison

Corpus documentation should identify the source, publication period, genre, sampling criteria, and preprocessing procedures for each collection. Texts should be selected

independently of whether they contain words or constructions that appear similar across the two languages.

Word counts alone may not provide a sufficient basis for comparison because morphological segmentation differs across languages. Both orthographic token counts and relevant morphological measures should therefore be documented (Kim, 2021).

3.3. Data Preprocessing

Tamil and Korean require language-specific preprocessing procedures.

For Tamil, the original Tamil script should be preserved alongside standardized transliterations and, where appropriate, International Phonetic Alphabet representations. Phonological analysis must account for relevant consonant distinctions, vowel length, and morphophonological alternations.

For Korean, the original Hangul representation should be preserved while spacing, punctuation, lexical stems, and grammatical particles are treated consistently.

The data structure should distinguish the following elements:

- Original orthographic form.
- Standardized transliteration.
- Phonological representation.
- Lemma or lexical stem.
- Morphological segmentation.
- Grammatical annotation.
- Semantic interpretation and source documentation.

Separating these representations helps prevent orthographic differences from being confused with phonological differences and makes it possible to compare equivalent linguistic units (Charles, 2025).

3.4. Lexical Similarity Measurement

Normalized Levenshtein similarity provides one measure of formal resemblance between two lexical representations. The score is calculated as follows:

$$S(a, b) = 1 - \frac{d(a, b)}{\max(|a|, |b|)}$$

Here, $d(a, b)$ is the minimum number of insertions, deletions, and substitutions required to transform one string into another. The denominator represents the length of the longer string.

The score ranges from 0 to 1, with higher values indicating greater formal similarity.

For instance, the strings KAR and KARU differ by one insertion. Their normalized similarity is:

$$S = 1 - \frac{1}{4} = 0.75$$

This calculation illustrates the mathematical interpretation of normalized edit distance. In linguistic analysis, however, the result must be interpreted in relation to the representation being compared.

The primary comparison should use standardized phonological representations wherever reliable data are available. Analyses using alternative transliteration or normalization procedures can help determine whether the results depend on transcription choices.

Lexical similarity scores should not be interpreted as probabilities of common ancestry. Semantic correspondence, regular phonological patterns, borrowing history, and independent etymological evidence must be considered separately (

3.5. Morphological and Syntactic Coding

A predefined codebook provides a consistent framework for annotating the grammatical characteristics of Tamil and Korean.

Table 4. Coding Framework for the Comparative Corpus Analysis of Tamil and Korean

Code	Category	Coding rule	Evidence examined
LEX	Lexical similarity	Compare corresponding	Basic vocabulary

		lexical concepts and phonological forms	
MOR	Morphological structure	Identify stems and grammatical morphemes	Inflected forms
SOV	Constituent order	Record the relative order of subjects, objects, and verbs	Transitive clauses
CAS	Case marking	Identify forms expressing grammatical case relations	Noun phrases
NEG	Negation	Record negative forms and their syntactic distribution	Negative clauses
TNS	Tense and aspect	Identify the expression of temporal and aspectual meanings	Verbal constructions
HON	Honorifics and politeness	Classify grammatical and lexical expressions of social relations	Honorific constructions
SEM	Semantic correspondence	Evaluate whether compared forms express equivalent or related meanings	Lexical entries and aligned sentences
BOR	Possible borrowing	Record independently supported evidence of borrowing	Historical and etymological sources
UNC	Uncertain analysis	Identify ambiguous or disputed linguistic interpretations	Uncertain segmentation or meaning

Coding should distinguish a grammatical function from the form used to express it. For example, a case suffix in Tamil and a Korean case particle may perform comparable functions without being formally or historically equivalent.

Independent annotation of a subset of the data can be used to assess coding consistency. Cohen's kappa is suitable for certain categorical annotation tasks, provided its assumptions are met. Disagreements should be documented and resolved through clearly specified coding rules.

3.6. Control Languages and Statistical Analysis

Control languages provide a baseline for evaluating whether observed similarities between Tamil and Korean are distinctive or widespread. For Tamil, relevant comparisons include other Dravidian languages, such as Telugu and Malayalam. Additional languages can be selected according to the feature under investigation.

Control selection should be feature-specific. Closely related languages are useful for identifying characteristics shared within a language family, whereas typologically different languages can help establish how widely particular grammatical patterns occur.

The quantitative analysis should include:

- Descriptive statistics for lexical similarity scores.
- Similarity distributions across predefined semantic categories.
- Normalized frequencies of selected grammatical constructions.
- Comparisons with control-language baselines.
- Confidence intervals and appropriate inferential tests.
- Sensitivity analyses for tokenization, word length, and transcription choices.

Where statistical comparisons are conducted, the analysis should account for dependence between observations and differences in corpus composition. Permutation procedures or matched comparisons may be appropriate when their assumptions are satisfied (Bhuvana, 2025).

Statistical significance should be interpreted alongside linguistic relevance. A measurable difference between similarity distributions does not, by itself, establish a historical relationship.

Chapter 4. Comparative Analysis and Findings

4.1. Lexical Correspondence

Lexical comparison requires the examination of both phonological form and semantic correspondence. Several Tamil and Korean lexical items illustrate why surface resemblance and lexical equivalence must be evaluated separately.

Table 5. Comparative Analysis of Basic Vocabulary in Tamil and Korean

Meaning	Tamil	Korean	Comparative interpretation
I	நான் (nāṇ)	나 (na)	Some superficial phonetic resemblance, insufficient to establish historical relatedness
You	நீ (nī)	너 (neo)	Partial surface resemblance; historical correspondence requires additional evidence
Rice/meal	சோறு (cōru)	밥 (bap)	Different lexical forms associated with broadly comparable meanings
Water	தண்ணீர் (taṇṇīr)	물 (mul)	Different lexical forms
Tree	மரம் (maram)	나무 (namu)	Different lexical forms
To eat	சாப்பிடு (cāppitu)	먹다 (meokda)	Different lexical forms and morphological conventions

The comparison reveals that words expressing comparable meanings can differ substantially in their phonological forms. Conversely, a partial resemblance between two forms does not establish that they are cognates.

Short lexical items require particular caution because their limited sound sequences increase the possibility of accidental resemblance. A rigorous lexical analysis must therefore compare predefined vocabulary categories, document the criteria used to identify potential correspondences, and examine alternative explanations.

Semantic equivalence also requires careful treatment. Words translated using the same English gloss may differ in cultural associations, conventional usage, or semantic range. The English translation should therefore function as a reference point rather than as the sole criterion for establishing lexical equivalence (Balaji & Jp, 2023).

4.2. Syntactic Structure and Constituent Order

Tamil and Korean both frequently exhibit SOV ordering in ordinary declarative clauses. The sentences *nāṇ cōru cāppittēṇ* and *naneun babeul meogeotda* illustrate this arrangement: the subject precedes the object, and the verb occurs at the end.

Nevertheless, sentence-level similarities must be distinguished from the complete grammatical systems of the two languages.

Table 6. Comparative Analysis of Syntactic Structures and Information Organization in Tamil and Korean

Feature	Tamil	Korean	Comparative finding
Basic constituent order	Predominantly SOV	Predominantly SOV	Broad typological similarity
Subject expression	Nominative subjects and contextually omitted subjects	Case-marked or topic-marked subjects and contextually omitted subjects	Different marking systems and discourse conditions
Object expression	Case marking and context-sensitive constructions	Case particles and context-sensitive constructions	Comparable functions with different forms
Verb placement	Commonly clause-final	Commonly clause-final	Similar basic ordering tendency
Information structure	Topic and focus expressed through grammatical and discourse resources	Topic and focus expressed through particles and discourse resources	Requires analysis beyond constituent order

The shared SOV tendency is a meaningful typological observation. However, SOV ordering is also found in many other languages, so it does not independently demonstrate a special historical connection between Tamil and Korean.

A corpus-based syntactic analysis should examine the frequency of SOV ordering across comparable clause types, including transitive clauses, subordinate clauses, and constructions involving topicalization. It should also distinguish canonical order from alternative orders used for discourse or pragmatic purposes (Bhuvana & Amala Karthiga, 2026).

4.3. Morphological Organization and Case Marking

Both languages employ grammatical marking to express relationships between lexical elements, but their morphological systems differ in their specific forms and distributions.

Tamil uses nominal case suffixes and verbal inflection. Korean employs postnominal particles and verbal endings. Although these systems can express comparable grammatical functions, the markers should not be treated as direct historical equivalents without additional evidence.

For example, Korean *namureul* (나무를, “tree” with accusative marking) contains the noun *namu* and the object particle *reul*. Tamil *marattai* (மரத்தை, “tree” with accusative marking) expresses the corresponding object function through Tamil morphology. The two forms are functionally comparable in suitable contexts, but their phonological structures and morphological histories differ.

This comparison demonstrates the importance of separating grammatical function from surface form. A shared function does not necessarily imply a shared morpheme or common historical origin (Bhuvana & Karthiga, 2025).

A rigorous analysis should examine the distribution of case marking across different noun types, syntactic environments, and discourse contexts. It should also account for differences in optionality and the conditions under which grammatical relationships can be recovered from context.

4.4. Verbal Morphology, Negation, and Honorifics

Verbal morphology provides another important area of comparison. Tamil verbal forms can encode tense, agreement, mood, and other grammatical distinctions. Korean verbal endings express tense, mood, speech level, and related grammatical or interpersonal meanings.

The two languages also possess multiple strategies for expressing negation. A meaningful comparison must consider the position of negative elements, their compatibility with different verb types, and their interaction with tense and other grammatical categories (Jou, 2024).

Honorifics and politeness provide a further area of comparison. Both languages employ linguistic resources for managing social relationships, but Korean has a particularly extensive system of speech levels and honorific marking. Tamil also expresses respect and social distinctions through grammatical and lexical choices, though the organization and distribution of these expressions differ.

Table 7. Comparative Analysis of Tense, Aspect, Negation, Honorifics, and Morphological Complexity in Tamil and Korean

Analytical dimension	Tamil	Korean	Research implication
Tense and aspect	Expressed through verbal morphology and constructional patterns	Expressed through verbal endings and constructional patterns	Compare grammatical categories and distributions
Negation	Multiple constructions depending on context	Multiple constructions depending on context	Compare negative constructions rather than isolated words
Honorifics	Socially conditioned grammatical and lexical forms	Honorific forms and speech-level distinctions	Separate communicative function from grammatical realization
Morphological complexity	Extensive suffixation and inflection	Extensive suffixation and particle use	Compare morphological organization at equivalent levels

The comparison supports a broader distinction between typological similarity and structural identity. Languages can share communicative functions while differing in how those functions are grammatically organized.

4.5. Interpretation of Linguistic Similarities

The comparative analysis identifies several dimensions along which Tamil and Korean can be meaningfully examined. These include constituent order, morphological marking, case relations, verbal morphology, and the linguistic expression of social relationships.

However, the evidence must be interpreted according to the type of similarity observed. SOV ordering represents a typological characteristic, while similarities between particular lexical forms require phonological and semantic investigation. Historical relationships demand a separate evidential standard (Jose et al., 2012).

The use of control languages is particularly important because some characteristics may be widespread across languages. If a pattern occurs in numerous unrelated languages, its presence in Tamil and Korean provides limited evidence of a distinctive relationship between them.

Lexical comparisons likewise require attention to word length, phonological representation, semantic range, and borrowing. Apparent similarities may become less pronounced when evaluated using standardized phonological representations or compared with appropriate baseline distributions (Sohn, 2001).

Taken together, these considerations support a multidimensional approach in which lexical, morphological, and syntactic observations are analyzed separately before their broader implications are considered (Bendale et al., 2024).

Chapter 5. Discussion and Conclusion

5.1. Discussion

The comparison of Tamil and Korean demonstrates the importance of treating linguistic similarity as a multidimensional phenomenon. Lexical resemblance, grammatical organization, and historical relatedness are distinct analytical questions, each requiring different forms of evidence.

Both languages exhibit predominantly SOV word order and substantial morphological marking. These characteristics provide a basis for meaningful typological comparison. Nevertheless, their lexical systems, grammatical particles, case marking, and verbal morphology differ in important respects.

Corpus linguistics provides a systematic approach to evaluating the distribution of these characteristics. Rather than relying exclusively on isolated expressions, researchers can investigate recurring patterns across predefined samples and compare their distributions across languages. Quantitative measures can identify similarities that warrant further examination, while qualitative analysis can explain their grammatical functions and limitations (Mokgokong, 2008).

The distinction between grammatical function and grammatical form is especially important. Tamil case suffixes and Korean case particles can perform comparable roles without being formally identical. Likewise, both languages can place verbs at the end of clauses without sharing an exclusive historical relationship on that basis.

Control-language comparisons strengthen the interpretation by identifying features that are

widespread rather than distinctive. This approach reduces the risk of overinterpreting common typological tendencies and provides a clearer basis for evaluating competing explanations.

5.2. Theoretical and Methodological Contributions

The study contributes to comparative linguistic research in three principal ways.

First, it integrates lexical comparison with morphological and syntactic analysis. This broader perspective prevents the investigation of linguistic similarity from being reduced to the identification of similar-looking words (Schultz, 2002).

Second, it establishes a distinction between surface resemblance, structural similarity, and historical relatedness. This distinction provides a more rigorous basis for interpreting cross-linguistic evidence and evaluating claims about linguistic affinity.

Third, it establishes a reproducible analytical framework incorporating corpus selection, lexical normalization, grammatical coding, and control-language comparison. The framework can be adapted to other language pairs, provided that the selection criteria and analytical assumptions are documented (Murasugi, 2018).

The principal methodological contribution is the integration of quantitative measurement with linguistic interpretation. Similarity scores become more informative when considered alongside grammatical structure, semantic correspondence, and historical evidence.

5.3. Limitations

Several methodological considerations affect the interpretation of cross-linguistic comparisons.

Differences in writing systems and transliteration conventions can influence lexical similarity measurements. Phonological representations are therefore preferable when the necessary linguistic information is available (Alagu & Priyalakshmi, 2023).

Morphological segmentation also differs between languages. Analyses based only on orthographic word forms may overlook meaningful grammatical correspondences or exaggerate differences between constructions (Sankaran, 2021).

Corpus composition introduces another potential limitation. Differences in genre, register,

and subject matter may affect lexical frequencies and syntactic distributions. Comparable sampling procedures and transparent documentation are necessary to minimize these effects.

Finally, typological and lexical similarities cannot independently establish historical relatedness. Such conclusions require regular sound correspondences, appropriate historical evidence, and careful consideration of borrowing and alternative explanations.

These limitations underscore the importance of evaluating linguistic patterns according to the strength and type of evidence available (Yeo et al., 2023).

5.4. Conclusion

This study examines linguistic similarities between Tamil and Korean through a comparative framework encompassing lexical correspondence, morphological organization, syntactic structure, and grammatical function. The comparison demonstrates that both languages share broad typological characteristics while differing in the specific forms and distributions of their linguistic elements (Arokiyaraj et al., 2021).

Their predominantly SOV word order and extensive morphological marking provide a basis for systematic comparison. However, these similarities are not sufficient to establish a common historical origin. Lexical resemblance must likewise be evaluated through phonological correspondence, semantic evidence, and historically grounded analysis (Li, 2023).

The study emphasizes that reliable linguistic comparison requires consistent data selection, transparent analytical procedures, and appropriate control languages. By distinguishing surface resemblance from structural similarity and historical relatedness, researchers can investigate language similarities without conflating descriptive patterns with claims about ancestry.

Future comparative research can strengthen this framework through systematic corpus analysis, validated grammatical annotation, and statistical evaluation of similarities across multiple language pairs. Such research would provide a more precise understanding of which characteristics are shared, how they are distributed, and what conclusions they can legitimately support (Kannan, 2011).

The broader contribution of this study lies in its emphasis on methodological rigor. Linguistic similarities are most informative when examined systematically, compared against

alternative explanations, and interpreted within the appropriate theoretical and historical context.

References

Abdullah, H. (2009). Relationship between national product and Malaysian government development expenditure: Wagner's law validity application. *International Journal of Business and Management*.

Arokiyaraj, S. (2020). India–Korea ancient maritime trade relations: Korean-Tamil language, food and culture similarities.

Arokiyaraj, S. (2026). Exploring Tamil–Korean linguistic parallels: A computational and historical analysis of possible Pre-Hangul contact. *Linguistics and Culture Review*.

Arokiyaraj, S., Ravichandran, G., Chozhan, A., & Narayanan, K. (2021). (Korean-Tamil) language and cultural similarities, maritime trade between early historic Tamilakam and Korea.

Balaji, V., & Jp, A. (2023). CSE_SPEECH@LT-EDI-2023: Automatic speech recognition vulnerable old-aged and transgender people in Tamil. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 204–208).

Bendale, A., Sapienza, M., Ripplinger, S., Gibbs, S., Lee, J., & Mistry, P. (2024). Sutra: Scalable multilingual language model architecture [Preprint]. arXiv. <https://arxiv.org/abs/2405.06694>

Bharath, A., & Madhvanath, S. (2007). Hidden Markov models for online handwritten Tamil word recognition. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* (Vol. 1, pp. 506–510). IEEE.

Bhuvana, V. (2025). A graphical comparison between Tamil and Korean languages: வரிவடிவ ஒப்பாய்வு நோக்கில் தமிழ் மற்றும் கொரிய மொழி. *Tamilmanam International Research Journal of Tamil Studies*, 1(5), 273–279.

Bhuvana, V., & Amala Karthiga, G. (2026). Hangul sound found in Tamil language. *Shanlax International Journal of Tamil Research*.

Bhuvana, V., & Karthiga, G. A. (2025). Cultural relations between Tamil Nadu and Korea: தமிழகமும் கொரியாவின் பண்பாட்டு உறவுகளும். *Tamilmanam International Research Journal of Tamil Studies*, 1(4), 208–220.

Charles, A. (2025). The incredible link between Tamil Nadu and Korea: Linguistic and cultural similarities between the communities. *Culture, Identity and Humanities*, 76.

Cho, S., & Park, H. (2024). Exploring the possibility of deciphering Tolkāppiyam through Korean grammar. *Barun Academy of History*.

Jou, E. (2024). Honorification as Agree in Korean and beyond. *Glossa: A Journal of General Linguistics*, 9(1).

Jose, J. M., Vu, N. T., & Schultz, T. (2012). Initial experiments with Tamil LVCSR. In *Proceedings of the International Conference on Asian Language Processing* (pp. 81–84). IEEE.

Kannan, N. (2011). Tamil–Korean relationship. In *Proceedings on the International Seminar on the Contributions of Tamils to the Composite Culture of Asia*.

Kim, J. (2021). A comparative study on Tamil and Korean language and customs: The similarity and its origin. *Shanlax International Journal of Tamil Research*.

Li, W. (2023). Morphologic, syntactic, and phonologic distance between Japanese and Altaic, Dravidian, Austronesian, and Korean languages. *European Journal of Theoretical and Applied Sciences*.

Mokgokong, P. (2008). Impact of a shrinking Arctic ice cover on marine primary production. *Geophysical Research Letters*.

Murasugi, K. (2018). Root infinitive analogues: Evidence from Tamil, Korean, and Japanese. In *Perspectives on the Architecture and Acquisition of Syntax: Essays in Honor of R. Amritavalli* (pp. 243–259). Springer Singapore.

Sankaran, L. (2021). “Talk in Tamil!”—Does Sri Lankan Tamil onward migration from Europe influence Tamil language maintenance in the UK? *International Journal of the Sociology of Language*, 2021(269), 123–149.

Schultz, T. (2002). GlobalPhone: A multilingual speech and text database developed at Karlsruhe University. In *Proceedings of Interspeech* (Vol. 2, pp. 345–348).

Shanmugamoorthy, M., Subbaiyan, A., Velusamy, S., & Mani, S. (2022). Review of groundwater analysis in various regions in Tamil Nadu, India. *KSCE Journal of Civil*

Engineering, 26(8), 3204–3215.

Sohn, H. M. (2001). *The Korean language*. Cambridge University Press.

Thanabalasingam, U. (2023). A phonetic comparison of Korean and Tamil. *Open Journal of Modern Linguistics*, 13(5), 711–733.

Alagu, V. R., & Priyalakshmi, G. (2023). Similarities between South Korea and Tamil Nadu: The unknown link. *Shanlax International Journal of Arts, Science and Humanities*.

Vasuthavan, N., & Sivanadhan, I. (2025). The influence of Tamil language on other languages. *International Journal of Academic Research in Business and Social Sciences*.

Yeo, E. J., Choi, K., Kim, S., & Chung, M. (2022a). Cross-lingual dysarthria severity classification for English, Korean, and Tamil. In *Proceedings of the 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 566–574). IEEE.

Yeo, E. J., Choi, K., Kim, S., & Chung, M. (2023). Speech intelligibility assessment of dysarthric speech by using goodness of pronunciation with uncertainty quantification [Preprint]. arXiv. <https://arxiv.org/abs/2305.18392>

Yeo, E. J., Kim, S., & Chung, M. (2022b). Multilingual analysis of intelligibility classification using English, Korean, and Tamil dysarthric speech datasets. In *Proceedings of the 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)* (pp. 1–6). IEEE.

Conflicts of interest

The author declares no conflicts of interest related to this research.

Data availability statement

The data analyzed in this study were obtained from publicly accessible sources. Details of the data sources and analytical procedures are provided in the manuscript.

Declaration of AI use

Artificial intelligence tools were used to assist with language editing and manuscript preparation. The author reviewed the manuscript and takes responsibility for its content, accuracy, and integrity.

This preprint was submitted under the following conditions:

- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a [Creative Commons CC-BY](#) license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.