Anusandhanvallari
Vol 2024, No.1
March 2024
ISSN 2229-3388

Creating Error Corpora for Learning Tamil Language for Plus Two Students of Madurai District

Mr. V. Selvakumar[1]* Dr. K. Umaraj[2]*

[1]Ph.D Full - Time Research Scholar, Department of Linguistics, Madurai Kamaraj University, Madurai-21.

[2]Associate Professor and Head**,** Department of Linguistics, Madurai Kamaraj University, Madurai-21

Abstract

Language acquisition is a complex cognitive process wherein learner errors, rather than being mere obstacles, provide critical diagnostic insights into the linguistic challenges faced by students. This paper explores the crucial role of Error Analysis (EA) and the development of a specialized error corpus to systematically study these phenomena in students learning Tamil at the higher secondary level. By employing a data-driven approach known as Computer-aided Error Analysis (CEA), this research aims to transcend anecdotal observations, moving toward a quantitative and qualitative understanding of learner difficulties. The core of this study involves creating a comprehensive learner corpus from authentic data collected from Plus Two students in the Madurai district. This data, encompassing written essays, formal assessments, and observational notes, is electronically stored and meticulously annotated using a custom, three-dimensional error tagging system. This system allows for the precise classification of errors based on their linguistic domain, nature, and affected word category. Processed using Java, the corpus facilitates the efficient retrieval of error patterns, serving as a foundational resource for language teachers, curriculum developers, and applied linguists to design targeted pedagogical tools for Tamil language education.

Keywords: Error analysis, corpus linguistics, annotation, language acquisition, Tamil language learning, Computer-aided Error Analysis (CEA).

Introduction

The historical trajectory of applied linguistics has witnessed fluctuating perspectives regarding the interpretation and pedagogical value of learner errors. Historically, errors were viewed strictly as failures in learning that needed to be eradicated through rigorous drilling. However, contemporary applied linguistics recognizes errors as natural, inevitable, and highly informative by-products of the language acquisition process. They represent the learner's developing transitional competence or 'inter-language.' In the context of Tamil language education, particularly for Plus Two (12th grade) students in Tamil Nadu, the transition from spoken dialects (Koduntamil) to formal, academic written Tamil (Centamil) presents unique orthographic, morphological, and syntactic challenges. Traditional Error Analysis (EA) previously faced criticism for its perceived lack of precision and a limited scope that often failed to capture the full picture of a learner's linguistic competence. It was sometimes viewed as overly subjective or detached from the broader context of language use.

However, the integration of Corpus Linguistics has revolutionized this field. Modern methodologies allow researchers to process vast amounts of linguistic evidence. By analyzing a large collection of learner data—capturing both accurate usage and deviations—researchers can move beyond simply identifying errors to understanding the underlying cognitive processes. This study focuses on this modern, data-driven approach, aiming to create a highly specialized, error-tagged corpus for Plus Two students in the Madurai district. Developing this resource will highlight common mistakes, reveal the sources of these errors, and offer a comprehensive view of the learning journey, ultimately bridging the gap between theoretical linguistics and practical classroom pedagogy.

Literature Review

The intersection of computational technology and linguistics has birthed Computer-aided Error Analysis (CEA), marking a paradigm shift in educational research. The theoretical backing for this approach is robust within the linguistic community. Corder (1967) fundamentally shifted the perspective on errors, arguing that they are significant because they provide to the researcher evidence of how language is learned or acquired, and what strategies or procedures the learner is employing.

Furthermore, as Ringbom (1987) noted, error analysis is a vital "key to a better understanding of the process underlying L2-learning." Ellis (1994) affirmed its continued value as a "useful tool" for identifying pedagogical priorities. The advent of Learner Corpora, pioneered by researchers like Sylviane Granger (1998), introduced standardized, computerized databases of learner language. Creating a learner corpus and annotating it with errors is a time-intensive and meticulous task requiring researchers to carefully parse through varied linguistic outputs. However, the long-term benefits are substantial; an error-tagged corpus serves as a rich, unparalleled resource that allows for automated and detailed statistical analysis, which is critical for refining Computer-aided Language Learning (CALL) programs.

Methodology

The process of developing this Tamil error corpus is a multi-phase undertaking, meticulously designed to ensure the data is comprehensive, accurate, and highly applicable for future research and syllabus design.

1. Coverage and Demographics

The study focuses on the prescribed Tamil textbook for Plus Two students. The research is geographically and demographically centered on higher secondary students from the Madurai district of Tamil Nadu. Madurai was selected due to its historical significance as a hub of the Tamil language and its distinct regional dialect, which provides a rich context for observing the interference of spoken forms in formal academic writing.

2. Data Collection

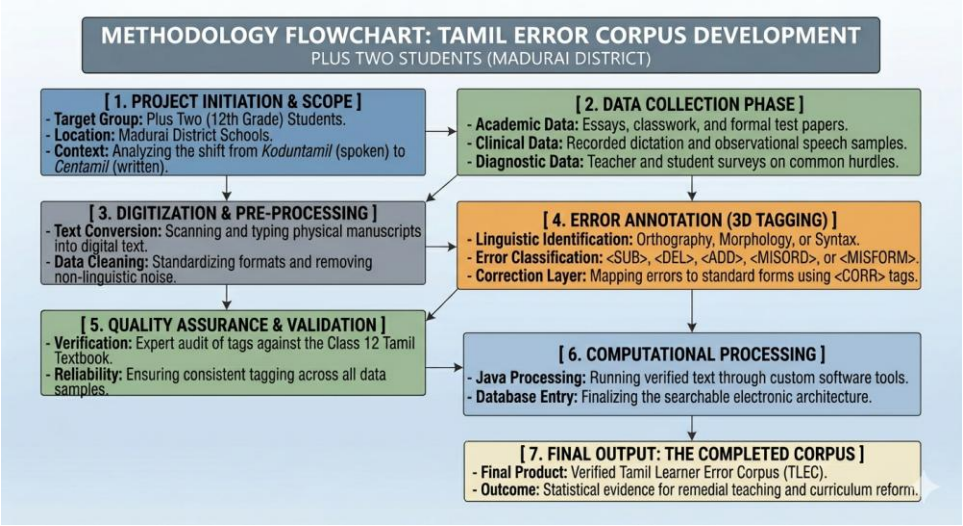
To capture a holistic view of learner competence, a triangulated approach was utilized to gather authentic learner data. The instruments include:

- Collected class test papers and term examinations around the Madurai district higher secondary schools to gauge controlled linguistic output under academic pressure.

- Direct observation of learners' speech and dictation exercises to capture spontaneous language use and phonetic-to-orthographic mapping errors.
- Specially designed diagnostic tools to pinpoint specific, reoccurring language problems identified by educators.

3. Data Analysis, Tagging, and Verification

The collected data is digitized and stored electronically. A systematic error tagging system is employed to annotate the corpus. This involves embedding specific XML-style codes into the text to mark and classify different types of deviations. The system is designed to be informative yet manageable, flexible for adding or removing tags, and consistent across different annotators. To ensure scientific rigor and inter-rater reliability, the entire tagged corpus undergoes a strict verification process.



4. Computational Processing

The entire dataset is processed and compiled using customized software tools developed in Java. This programming backbone ensures the corpus is easily searchable, allowing researchers to run queries based on specific error tags, word classes, or demographic variables, making it a dynamic tool for computational analysis.

The Error Tagging System And Sample Examples

A robust annotation framework is vital for the utility of the corpus. Our error tagging system utilizes a three-dimensional taxonomy, combining:

- Linguistic Category: (e.g., Orthography, Morphology, Syntax, Lexicon).
- Nature of the Error: (e.g., Substitution, Omission, Addition, Misordering).
- Target Word Class: (e.g., Noun, Verb, Case Marker).

This approach, advocated by James (1998), provides a comprehensive and descriptive classification without the complexities of subjectively interpreting the psychological source of the error. To facilitate this complex work, a custom, menu-driven editor was developed. This tool allows annotators to insert error tags (<TAG>) and corrected forms (<CORR>) by simply clicking on relevant options, streamlining a traditionally laborious task.

Below are five detailed examples extracted from the corpus, categorized according to the rules and standards from the Class 12 Tamil Textbook (2024 Edition):

Example 1: Substitution Error (SUB) – Orthographic/Sandhi Rules

- **Textbook Reference:** Chapter 1: "Inikkum Ilakkanam" (Avoiding Spelling Errors)
- **Student's sentence:** *நான் இயற்கை காட்ச்சி கண்டேன்.*
- **Intended meaning:** I saw the natural scenery.
- **Error:** The student added an unnecessary 'ச்' (ch) after 'ட்' (t). According to standard textbook rules, the consonants 'இட்' and 'இற்' should not be followed by their own hard consonant (Vallinam) counterparts in this specific word structure.
- **Annotated text:** *நான் இயற்கை* _{*காட்ச்சி*}<CORR>*காட்சி*</CORR> *கண்டேன்.*
- **Explanation:** The <SUB> tag identifies a substitution of a phonetically similar but orthographically incorrect cluster. This aligns with textbook instructions on avoiding 'ezhuthu pizhai' (spelling mistakes) in words containing specific consonant combinations.

Example 2: Omission Error (DEL) – Accusative Case Marker

- **Textbook Reference:** Chapter 1: "Thodariyal Pokkugal" (Syntactic Trends)
- **Student's sentence:** *கபிலன் பாடம் படித்தான்.*
- **Intended meaning:** Kapilan read the lesson.
- **Error:** In formal Tamil, when a noun acts as a definite object, the second case marker (Accusative '-ai') is required. The student omitted the marker, rendering the sentence grammatically incomplete for academic standards.
- **Annotated text:**
- *கபிலன்* <DEL/Case>*பாடம்*</DEL/Case><CORR>*பாடத்தைப்*</CORR> *படித்தான்.*
- **Explanation:** The <DEL/Case> tag signifies the deletion of a functional case marker. The curriculum emphasizes the use of 'vetrumai urubu' (case markers) to clearly distinguish the subject from the object.

Example 3: Addition Error (ADD) – Concord/Agreement

- **Textbook Reference:** Chapter 2: "Thodar Amaippu" (Sentence Structure)
- **Student's sentence:** *அவர்கள் நேற்று வந்தான்.*
- **Intended meaning:** They came yesterday.

- **Error:** The student used a plural subject (Avargal - They) alongside a singular masculine verb (vandhaan - He came). This represents a 'Thinai-Paal-En' (Category-Gender-Number) mismatch.
- **Annotated text:** அவர்கள் நேற்று <ADD>வந்தான்</ADD><CORR>வந்தார்கள்</CORR>.
- **Explanation:** The <ADD> tag highlights a failure in Subject-Verb agreement. The Plus Two syllabus heavily focuses on maintaining consistency between the subject's gender/number and the corresponding verb suffix.

Example 4: Misformation Error (MISFORM) – Semantic Confusion (La/La/La)

Textbook Reference: Chapter 1: "La-ka-ra, La-ka-ra, Zha-ka-ra Rules"

Student's sentence: தோட்டத்தில் வாலை மரம் உள்ளது.

Intended meaning: There is a banana tree in the garden.

Error: The student utilized 'வாலை' (Vaalai - tail) instead of the correct lexical item 'வாழை' (Vaazhai - banana).

Annotated text: தோட்டத்தில் <MISFORM>வாலை</MISFORM><CORR>வாழை</CORR> மரம் உள்ளது.

Explanation: The <MISFORM> tag addresses 'Porul Mayakkam' (semantic confusion). Distinguishing between the three lateral approximants ('ல', 'ள', 'ழ') is a primary orthographic challenge among students, explicitly addressed in the textbook exercises.

Example 5: Misordering Error (MISORD) – Punctuation and Meaning

- **Textbook Reference:** Chapter 3: 'kaarpulliyum porul mayakkamum' (Commas and Meaning)
- **Student's sentence:** அவள் அக்காள், வீட்டிற்குச் சென்றாள்.
- **Intended meaning:** She went to her elder sister's house.
- **Error:** The incorrect placement of the comma (kaarpulli) fundamentally alters the subject and meaning of the sentence. In the student's version, the "elder sister" is the one going home, rather than the intended subject going to the "sister's house."
- **Annotated text:**

<MISORD>அவள் அக்காள், வீட்டிற்குச் சென்றாள்</MISORD><CORR>அவள், அக்காள் வீட்டிற்குச் சென்றாள்</CORR>.

- **Explanation:** The <MISORD> tag indicates that while the vocabulary is correct, the syntactic arrangement or punctuation defining their relationship is flawed, affecting the 'thodar'.

Discussion

The initial findings from the annotated corpus reveal significant patterns in the linguistic development of Plus Two students. A major portion of the orthographic errors stems from the influence of spoken Tamil on written output. In spoken Tamil, distinct phonemes (like 'ல', 'ள', 'ழ' or 'ற', 'ர') are often conflated, leading to misformations

when students attempt to write in Centamil. Furthermore, syntactic errors indicate that while students understand the semantic meaning of their sentences, they frequently struggle with the rigid morpho-syntactic rules of formal Tamil, such as strict subject-verb agreement and mandatory case marking. By cataloging these specific pain points, the corpus provides empirical evidence that can be directly used to design targeted remedial materials.

Conclusion

The creation of a comprehensive, error-tagged corpus tailored for Tamil language learners represents a significant step forward in improving language teaching and learning methodologies. Historically, Error Analysis faced criticism for a perceived lack of precision. However, this study demonstrates that integrating modern corpus linguistics revitalizes the field, providing a robust, systematic, and highly quantitative approach to understanding the underlying processes of language learning. By embracing a data-driven methodology known as Computer-aided Error Analysis (CEA), this study has systematically outlined a multi-phase framework—spanning from authentic data collection in the Madurai district to rigorous Java-based computational analysis. The implementation of a custom, three-dimensional error tagging system allows for an unprecedented, precise classification of learner mistakes such as substitutions, omissions, additions, and misorderings.

By providing a detailed perspective on persistent learner difficulties, this finalized, verified corpus will serve as a foundational pedagogical resource. It will directly inform the creation of targeted teaching methodologies, empowering teachers, curriculum developers, and applied linguists to design more effective educational tools. Ultimately, this effort seeks to significantly enhance the quality of Tamil language education, assisting current and future generations of students in navigating the complexities of academic Tamil with greater ease and accuracy.

References

[1] Brown, H. D. (2000). Principles of language learning and teaching. New York: Longman.

[2] Corder, S. P. (1967). The significance of learners' errors. International Review of Applied Linguistics in Language Teaching, 5(4), 161-170.

[3] Corder, S. P. (1973). Introducing applied linguistics. Middlesex: Penguin.

[4] Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. System, 26(2), 163-174.

[5] Dulay, H., Burt, M., & Krashen, S. (1982). Language Two. New York: Oxford University Press.

[6] Ellis, R. (1994). The Study of Second Language Acquisition. Oxford: Oxford University Press.

[7] Ellis, R. (1997). Second language acquisition. Oxford: Oxford University Press.

[8] Granger, S. (Ed.). (1998). Learner English on computer. London & New York: Addison Wesley Longman.

[9] James, C. (1998). Errors in language learning and use: Exploring error analysis. London & New York: Longman.

[10] Keshavarz, M. H. (1997). Contrastive analysis and error analysis. Tehran: Rahnama Publications.

[11] Lengo, N. (1995). What is an error? English Teaching Forum, 33(3), 20-24.

[12] Nadaraja Pillai, N., & Vimala, S. (1981). Error Analysis. Mysore: Central Institute of Indian Languages (CIIL).