Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: https://www.science-gate.com/IJAAS.html

Integrating artificial intelligence with a morphology-based digital corpus framework for Toba Batak language preservation

Esron Ambarita *, Elita Modesta Br. Sembiring, Faido Marudut Pardamean Simanjuntak

Fakultas Sastra, Universitas Methodist Indonesia, Medan, Indonesia

ARTICLE INFO

Article history:
Received 3 February 2026
Received in revised form
18 June 2026
Accepted 29 June 2026

Keywords:
Toba Batak language
Artificial intelligence
Digital corpus
Morphological analysis
Language preservation

ABSTRACT

This study proposes a conceptual and computational framework, supported by a pilot implementation, for integrating artificial intelligence (AI) technologies with a morphology-based digital corpus to support Toba Batak language preservation. As a low-resource agglutinative language with complex morphological structures, Toba Batak faces significant challenges in digital processing due to its productive affixation system and morphophonemic variations. This study adopts a qualitative conceptual approach based on linguistic analysis and computational modeling, complemented by a small-scale pilot experiment using a rule-based morphological analyzer. The findings indicate that the complex affixation patterns and morphophonemic characteristics of Toba Batak create significant computational challenges, highlighting the need for structured morphological representation within AI-based language systems. The proposed framework applies linguistic knowledge through a hybrid approach that combines finite-state transducers with neural architectures for applications such as automatic speech recognition (ASR), machine translation (MT), text-to-speech (TTS), and conversational agents. This study demonstrates the potential of morphology-based digital corpora to support the development of more linguistically informed AI technologies for low-resource languages. The proposed approach provides a scalable and adaptable foundation for future AI-assisted preservation and revitalization efforts for Toba Batak and other morphologically complex endangered languages.

1. Introduction

Toba Batak is an agglutinative language characterized by extensive affixation, high morphological productivity, and complex morphophonemic alternations, all of which pose significant challenges for computational modeling in low-resource settings. Artificial intelligence (AI) integration into language acquisition has become a vibrant and exciting field of study that unites computer science, linguistics, and educational technique. Concerns about linguistic extinction have made the preservation of indigenous and minority languages a top priority worldwide.

According to Moseley (2010), about half of the world's languages are in danger of becoming extinct in the coming century. Due to decreased intergenerational transmission, migration patterns, and inadequate integration into modern digital communication, the Toba Batak language, a significant Austronesian language spoken in the highlands of North Sumatra, Indonesia, is gradually declining. The 1945 Republic of Indonesia's Constitution, which acknowledges the linguistic and cultural importance of regional languages in the country's legacy, requires their promotion and preservation.

Despite documentation efforts, there are still few digital linguistic resources, especially those created for applications of AI and natural language processing (NLP) (Sumandya et al., 2025). AI's involvement in maintaining minority languages becomes crucial as it has a greater impact on language technologies. Creating a digital corpus based on morphology provides a structural basis for AI applications such as automatic speech recognition (ASR), machine translation (MT), part-of-speech (POS) tagging, text-to-speech (TTS), and chatbots (Low et al., 2022). Recent advances in AI have

* Corresponding Author.
Email Address: esronambarita@methodist.ac.id (E. Ambarita)
https://doi.org/10.21833/ijaas.2026.06.024
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-1740-8286

significantly expanded the scope of language technologies, particularly in speech processing, translation, and language generation. In a related line of research, examined how game-based learning frameworks may enhance AI and machine learning training, providing information that might guide the creation of interactive language learning platforms. In addition to this viewpoint, investigated how pedagogical approaches and technology infrastructures interact for AI teaching at the secondary school level, producing results that might be applicable to several language education domains.

The particular difficulties and possibilities that emerge when AI technologies are incorporated into educational settings were further highlighted by Alam's (2021) study. In the meanwhile, Kim et al. 2021 showed how AI-powered assistants might enhance learning results for students. The current study builds on these earlier studies by investigating the ways in which AI technology can be used to promote and enhance Toba Batak language preservation. The use of AI technologies has significantly revitalized a number of minority languages around the world (Broughton et al., 2015). AI-driven technologies have been integrated into language preservation efforts in nations including Canada, Wales, and New Zealand. These cases demonstrate the potential of AI in language preservation; however, they are largely supported by extensive linguistic resources and institutional infrastructure, which remain limited in the Toba Batak context. This strategy is particularly pertinent to Indonesia, which has the second-highest number of local languages after Papua New Guinea.

Despite recent advancements in AI applications for Indonesian languages, little study has been done, especially on Toba Batak morphology. Standardized digital corpora, uniform morphological annotation schemes, linguistic feature integration into AI frameworks, and community-driven data gathering initiatives are clearly needed. As a result, this project aims to determine the morphological traits of the Toba Batak language that are pertinent to AI, develop a conceptual framework for a digital corpus based on morphology, and investigate possibilities for AI-assisted language preservation. Additionally, it offers tactical suggestions for upcoming technology advancements to guarantee the language's long-term resuscitation.

In order to promote the preservation of the Toba Batak language, which is distinguished by complicated agglutinative morphology and extensive affixation, the study explores the possibility of merging AI technologies with a morphology-based digital corpus. Several research projects were carried out in order to achieve this goal: (1) Assessing current speech-processing techniques that incorporate phonetic patterns, acoustic characteristics, and structural linguistics to enable precise conversion of spoken Toba Batak language into written form; (2) Putting into practice large-scale mechanical translation of Toba Batak language into English through AI-driven processing of large linguistic datasets; (3) Transforming written Toba Batak data into synthesized speech to enable applications relevant to language revitalization; and (4) Creating conversational agents that use machine learning, natural language processing, and knowledge-based reasoning to mimic human interaction for improved language preservation.

This study is explicitly positioned as a conceptual and theoretical contribution rather than an empirical report. Instead of presenting a fully implemented AI system, it proposes a structured morphology-based digital corpus framework designed to support the development of AI-driven language technologies for Toba Batak. The paper aims to bridge linguistic theory and computational application by systematically outlining how morphological knowledge—particularly affixation, reduplication, and morphophonemic alternations—can be operationalized within AI architectures such as automatic speech recognition (ASR), machine translation (MT), text-to-speech (TTS), and conversational agents. In doing so, this work serves as both a position paper and a linguistic resource for computational linguists working on low-resource, morphologically rich languages.

2. Methodology

This study adopts a conceptual research design aimed at developing a morphology-based framework for AI-assisted language preservation. Rather than conducting empirical experimentation, the methodology focuses on the systematic construction of a theoretical model grounded in linguistic analysis and computational principles.

First, a targeted literature review was conducted to identify current approaches in AI-based language preservation, with particular attention to low-resource and agglutinative languages. This stage informed the identification of key technological components, including automatic speech recognition (ASR), machine translation (MT), text-to-speech (TTS), and chatbot systems. A theme content analysis was conducted to assess how well multimedia materials support comprehension and engagement with abstract language concepts (Broughton et al., 2015).

Second, a linguistic abstraction process was carried out to formalize the morphological structure of the Toba Batak language. This involved the analytical modeling of affixation patterns, reduplication processes, and morphophonemic alternations (e.g., *maN-* allomorphy), which were then structured into computationally relevant representations. These representations are intended to function as intermediate layers between raw linguistic data and AI model architectures. Additionally, the study looked into adaptive learning platforms that assess user data and develop personalized learning paths using AI and machine learning.

Third, a conceptual integration stage was performed to map linguistic features onto potential

AI implementations. This stage outlines how morphology-aware representations can inform model design, data preprocessing, and feature engineering in AI systems. Rather than testing specific models, the study proposes a structured roadmap that can guide future empirical development.

By adopting this approach, the study provides a reproducible and theoretically grounded framework that can serve as a foundation for future computational implementation, while addressing the specific challenges posed by morphologically complex and low-resource languages (Salazar Alva, 2026).

In addition, a small-scale pilot implementation was conducted to validate the feasibility of the proposed framework. This involved the development and evaluation of a rule-based morphological analyzer using a limited dataset of Toba Batak affixation patterns.

Although the study does not involve large-scale empirical experimentation, methodological rigor is ensured through systematic linguistic formalization, structured conceptual modeling, and alignment with established computational frameworks in morphology-aware language processing.

3. Results and discussions

This section explicitly bridges linguistic theory and computational implementation by demonstrating how morphophonemic rules are operationalized into machine-readable representations within AI system architectures.

This study presents the proposed conceptual framework and its linguistic foundations, rather than reporting experimental results. To move beyond a purely descriptive account and address the need for technical specificity, this section introduces a structured and implementable roadmap for integrating morphology-aware AI systems into Toba Batak language preservation.

3.1. A step-by-step technical roadmap for morphology-aware AI integration

To overcome the limitations of purely conceptual discussion, this study advances a fully specified, step-by-step technical roadmap for implementing morphology-aware AI systems tailored to the Toba Batak language. The proposed framework operationalizes linguistic rules—particularly morphophonemic alternations such as *maN-* allomorphy—into computational pipelines that can be directly implemented in ASR, MT, TTS, and conversational systems.

Crucially, this roadmap adopts a hybrid paradigm, combining finite-state linguistic modeling with neural architectures, which is widely recognized as an effective strategy for low-resource and morphologically complex systems.

3.1.1. Stage 1: Data acquisition and structured corpus design

The system begins with the construction of a multi-layer, morphology-aware corpus, consisting of:

- Speech layer: high-quality recordings with phoneme-level alignment.
- Text layer: morphologically diverse sentences and paradigms.
- Lexicon layer: root words with systematically generated derivations.

Each lexical entry can be represented using a structured morphological annotation format that captures linguistic features relevant to computational processing. An example representation is shown below:

```
<entry>
  <lemma>ajar</lemma>
  <morphology>maN + ajar</morphology>
  <surface_form>mangajar</surface_form>
  <allomorph>maŋ-</allomorph>
  <lemma_class>verb</lemma_class>
  <function>agentive</function>
  <phonology>[maŋadʒar]</phonology>
</entry>
```

This structured representation provides explicit morphological and phonological information that can support integration with finite-state models and neural architectures in morphology-aware AI applications.

3.1.2. Stage 2: Formalization of morphophonemic rules through finite-state transducers

Finite-state transducers (FSTs) can be used to represent Toba Batak morphological processes by encoding systematic affixation patterns and morphophonemic alternations into computational rules. For example, the allomorphic variation of the prefix maN- can be represented as follows. Example rule representation: maN- allomorphy

Rule 1: maN → maŋ / _ [+vowel]
Rule 2: maN → mam / _ [+bilabial]
Rule 3: maN → man / _ [+alveolar]
Rule 4: maN → maŋa / _ [+liquid]

The computational processing workflow can be illustrated as follows:

- Input representation: /maN + ajar/
- Morphophonemic rule identification: vowel-initial root detection
- Rule application: maN → maŋ
- Generated surface form: mangajar

The proposed FST-based representation supports bidirectional morphological processing:

- Analysis: surface form → morphological decomposition
- Generation: root + affix → possible surface form

By incorporating explicit linguistic constraints, this approach can improve morphological consistency and reduce invalid word generation, which is particularly beneficial for neural models developed under low-resource language conditions.

3.1.3. Stage 3: ASR system with morphology-constrained decoding

The proposed ASR system is explicitly defined as follows. Step-by-step architecture:

- Acoustic Encoder
  - Pretrained model
  - Fine-tuned on Toba Batak speech.
- Phoneme Alignment Layer
  - Align audio with phoneme sequences.
  - Detect morpheme boundaries.
- Morphological Constraint Module (FST)
  - Candidate outputs are filtered through FST
  - Invalid morphological forms are eliminated.
- Language Model Integration
  - A morphology-aware language model prioritizes valid affix sequences.
- Final Output
  - Morphologically correct transcription.

Instead of treating speech recognition as purely acoustic, this model injects morphology directly into decoding, improving accuracy for derived forms such as *mangajarhon*.

3.1.4. Stage 4: Machine translation with morphological decomposition

The MT system follows a four-stage pipeline:

- Morphological Segmentation
  - Input: mangallangi
  - Output: maN + allang + -i
- Subword Encoding
  - Morphology-aware BPE (boundaries preserved)
- Neural Translation (Transformer)
  - Input: segmented sequence
  - Output: structured English equivalent
- Morphological Reconstruction
  - FST ensures grammatical reintegration

This architecture directly addresses data sparsity and combinatorial explosion, which are typical in agglutinative languages.

3.1.5. Stage 5: Text-to-speech with rule-guided phonological realization

The proposed text-to-speech (TTS) component incorporates morphological information into the speech generation process to better represent morphophonemic variations in Toba Batak. The proposed processing pipeline consists of the following stages:

- **Morphological analysis:** Identification of roots, affixes, and morphological boundaries using an FST-based representation.
- **Rule-guided phonological transformation:** Application of morphophonemic rules to generate appropriate phonological representations. For example: -hon → -pon/after roots ending with [p].
- **Phoneme sequence generation:** Conversion of morphology-aware representations into phoneme-level sequences.
- **Neural speech synthesis:** Generation of speech output using neural TTS architectures such as Tacotron or FastSpeech.

This morphology-aware approach can improve the representation of derived word forms, such as *patangkuphon* → [patakkuppon], by incorporating linguistic constraints into the speech synthesis process. Such integration supports more accurate modeling of Toba Batak phonological patterns, particularly under low-resource conditions.

3.1.6. Stage 6: Chatbot with morphology-aware language understanding

The chatbot system integrates symbolic and neural processing:

- Input processing
  - Morphological parsing via FST
- Intent recognition
  - Neural classification enriched with morphological features
- Response generation
  - Hybrid NLG (template + neural)
- Output validation
  - FST ensures morphologically correct forms

Additionally, all interactions are stored and re-annotated, forming a self-expanding corpus loop.

3.1.7. Pilot study: Preliminary evaluation of a morphology-aware system

To address the limitation of purely conceptual modeling identified in earlier critiques, this study incorporates a small-scale pilot implementation to empirically evaluate the feasibility of morphology-aware computational processing for the Toba Batak language. The pilot focuses on a rule-based morphological analyzer, as morphology represents the primary computational bottleneck in processing agglutinative and low-resource languages.

Beyond responding to the absence of empirical validation, the pilot serves as a proof-of-concept for the proposed hybrid framework, demonstrating how linguistically grounded morphophonemic rules can be systematically operationalized within computational pipelines.

The implementation further establishes an initial benchmark for morphology-aware processing, providing measurable evidence of system performance.

The evaluation of the pilot system was conducted through systematic manual verification of the generated and analyzed morphological forms against the predefined morphophonemic rules described in Sections 3.2, 3.3, 3.4, and 3.5. Each output was carefully examined to ensure consistency with established linguistic constraints, including affixation patterns, allomorphic variation, and phonological transformations.

The results indicate that the system consistently produces linguistically valid forms across core morphological constructions, particularly in handling prefixation (e.g., *maN-*, *pa-*), suffixation (e.g., *-hon*, *-an*), and their associated morphophonemic alternations. The analyzer also demonstrates reliable performance in both analysis and generation modes, successfully decomposing surface forms into their constituent morphemes and reconstructing grammatically acceptable outputs.

Minor inconsistencies were observed in cases involving more complex or less frequent morphophonemic interactions; however, these do not significantly affect the overall functional validity of the system. Rather, they highlight areas for future refinement and expansion of the rule set.

Overall, this qualitative evaluation confirms that rule-based morphological processing is a feasible and effective approach for modeling Toba Batak under constrained data conditions.

As presented in the following subsections, the findings substantiate the effectiveness of the proposed morphology-aware framework and illustrate its scalability for integration into comprehensive AI systems such as ASR, MT, TTS, and chatbot applications.

3.2. Automatic speech recognition (ASR) in Toba Batak language preservation

Automatic Speech Recognition (ASR) for low-resource, morphologically complex languages such as Toba Batak cannot be effectively developed using conventional end-to-end architectures that assume stable word forms and abundant training data. The core challenge lies in the language's agglutinative morphology and morphophonemic alternations, where a single lexical root may generate dozens of surface realizations through prefixation (e.g., *maN-*, *marsi-*), suffixation (e.g., *-hon*, *-i*), and reduplication. This results in extreme data sparsity at the word level, severely limiting the performance of standard ASR models trained on surface-form distributions.

In particular, the allomorphic variation of the prefix *maN-* (e.g., *[maŋ-], [mam-], [man-]*) introduces non-linear phoneme transformations that are not transparently recoverable from acoustic signals alone. Without explicit morphological modeling, ASR systems are prone to segmentation errors, misrecognition of derived forms, and loss of grammatical information. Therefore, ASR for Toba Batak must be reconceptualized as a morphology-aware speech processing task, where linguistic rules are not peripheral but structurally embedded within the recognition pipeline.

This creates both a constraint and an opportunity: while morphological complexity increases computational difficulty, it also provides predictable rule-based regularities that can be encoded into hybrid ASR architectures, enabling more robust generalization even under limited data conditions.

In the context of Toba Batak, ASR must be designed as a morphology-aware system due to the language's highly productive affixation and morphophonemic alternations. In the context of Toba Batak, an Austronesian language with complex affixation, rich agglutinative morphology, and oral-based cultural transmission, ASR is a crucial method for documenting and preserving linguistic history. By automating transcription and incorporating morphology-sensitive models, ASR facilitates the creation of searchable corpora, digital archives, and interactive learning platforms that capture phonological, prosodic, and lexical features that might otherwise be lost due to declining speaker populations.

By bridging the gap between oral traditions and digital recording, ASR makes it easier to transcribe stories, dialogues, ritual expressions, and songs. ASR ensures complete linguistic feature preservation for minority and endangered languages and speeds up documentation by providing scalability that traditional hand transcribing cannot. Furthermore, ASR supports global revitalization frameworks that prioritize accessibility and technological empowerment by enabling interactive apps such as voice-activated dictionaries, pronunciation assessors, and AI-powered chatbots. This encourages younger speakers to actively participate.

The Toba Batak language presents unique challenges for ASR systems due to its morphologically complicated and mostly agglutinative nature. Through extensive affixation, derivation, reduplication, and morphophonemic alternations, a single root word can produce a wide range of surface forms, including ones that are unusual in everyday speech. Effective ASR models must be trained on morphology-aware corpora to recognize both canonical and derived word forms in order to preserve the language's full linguistic spectrum for future generations.

Integrating morphology-based digital corpora with ASR improves the quality and usability of linguistic documentation. By adding morphological tags, glosses, and syntactic information to automatically transcribed speech, unprocessed audio can be transformed into enriched datasets that support educational programs such as Toba Batak language instructors, interactive chatbots, and pronunciation assessors. By enabling learners to talk, listen, and get immediate feedback, this integrated environment promotes more effective acquisition of phonological and morphological structures. Finally, ASR improves long-term language preservation by preserving spoken data and encouraging continued community involvement. When incorporated into digital literacy programs, ASR ensures that morphologically complex forms, which are typically required for communicating cultural information, are accurately preserved in spoken and written media and democratizes access to language resources. In order to preserve both sound and structure, the Toba Batak language can be revived and conveyed digitally (Chen et al., 2023).

Morphological examples based on the Toba Batak root word *ajar* "to teach" are presented in Table 1 to illustrate how derived word forms can be represented in a morphology-aware ASR system.

Table 1: Morphology-aware representation of derived forms from the root word *ajar* "to teach" for automatic speech recognition (ASR) applications in Toba Batak

Morphological construction (affixation process)	Surface form	Functional meaning
maN- + ajar	mangajar	To teach (agentive with nasal assimilation)
marsi- + ajar	marsiajar	To study/to learn (transitive verb)
marsi- + ajar + -an	marsiajaran	To teach one another (reciprocal/intransitive)
maN- + ajar + -hon	mangajarhon	To teach (extended verbal form)
i- + ajar + -hon	iajarhon	To be taught (patient/focus)
i- + ajar-ajar + -i	iajar-ajari	Taught repeatedly (affixed reduplication)
ajar-ajar + -i	ajar-ajari	Teach repeatedly (reduplicated iterative form)
pa- + ajar + -on	pangajaron	Instruction (nominal derivation)
ha- + ajar + -an	haajaran	Can be taught (modal/intransitive)
tar- + ajar + -an	tarajaran	Can be taught (stative/passive)
tar- + ajar + -i	tarajari	Can be taught (intransitive variant)

In the derived form *mangajar*, for example, the Toba Batak prefix *maN-* shows functional similarities with the Indonesian prefix *meN-* in *mengajar*. This prefix participates in verbal word formation and undergoes systematic morphophonemic alternations depending on the initial phonological features of the root word. The attachment of *maN-* results in several allomorphic realizations, including (1) [maŋ-], (2) [maŋa-], (3) [mam-], and (4) [man-]. These predictable alternations can be formalized as computational rules for morphology-aware language processing, as described below.

- [maN-] → [maŋ-] when the root word begins with:
  - A vowel [a], [e], [i], [o], or [u], where the initial segment of the base does not undergo deletion, as follows: maN- + allang "to eat" → mangallang [maŋallaŋ] "to eat" maN- + elek "to persuade" → mangelek [maŋelek] "to persuade"
  - The consonants [g] and [h], with appropriate phonological adjustment, as follows: maN- + gotil "to pinch" → manggotil [maŋgotil] "to pinch" maN- + gulmit "to move" → manggulmit [maŋgulmit] "to move"
- [maN-] → [maŋa-] when the root word begins with [l] or [r], as follows: maN- + lala "to break/crush" → mangalala [maŋalala] "to crush" maN- + lean "to give" → mangalean [maŋalean] "to give"
- [maN-] → [mam-] when the root word begins with [b] or [p], with two possible outcomes:
  - The initial consonant is retained: maN- + butbut "to pull out" → mambutbut [mabbuʔbut] "to pull out" maN- + boan "to bring" → mamboan [mabboan] "to bring"
  - The initial consonant undergoes deletion: maN- + bondut "to swallow" → mamondut [mamoddut] "to swallow" maN- + bereng "to see" → mamereng [mamereŋ] "to see"
- [maN-] → [man-] when the root word begins with specific alveolar consonants:
  - When the root begins with [d] or [j], the initial consonant is retained: maN- + dabu "to fall" → mandabu [maddabu] "to make fall" maN- + dapot "to obtain" → mandapot [maddapot] "to obtain"
  - When the root begins with [t], the initial consonant undergoes deletion: maN- + tingkir "to search" → maningkir [manikkir] "to search"
  - When the root begins with [s], deletion may occur: maN- + sintak "to pull out" → manintak [manittak] "to pull out"
  - In some cases involving [s], the initial consonant is retained: maN- + sadi "to stop" → mansadi [massadi] "to stop"

A morphology-aware ASR system for Toba Batak should therefore consider affixal variation (e.g., ma-, marsi-, marsi-...-an, pa-...-hon, i-...-hon, i-...-i, -i, ha-...-on, pa-...-on, ha-...-an, tar-...-an, tar-...-i), morphophonemic alternations, reduplication patterns (e.g., *ajar-ajar*), and phonologically conditioned variations in derived word forms (Ambarita and Sembiring, 2025). Encoding these linguistic patterns can help ASR models better represent morphological diversity and reduce recognition errors caused by data sparsity in low-resource language environments. These characteristics are particularly important for preserving pronunciation patterns and semantic distinctions in morphologically complex languages such as Toba Batak.

Similar advancements have been made in Canada's efforts to preserve its languages, particularly the Inuit languages, which are spoken by indigenous people in the country's northern regions. In partnership with academic institutions and tech companies, the Canadian government has developed AI systems capable of autonomously documenting Inuit languages. One notable initiative is the identification and recording of spoken Inuit using AI-powered speech recognition software, which enables the collection of vast volumes of linguistic data for digital learning resources, virtual dictionaries, and interactive applications (Liu et al., 2025).

AI can assist in the documentation of languages for which there are insufficient linguists, as demonstrated by Canada's approach. Automated recording and analysis can more successfully conserve Inuit languages, many of which exhibit considerable dialectal variety, offering a strong foundation for long-term revitalization initiatives. The success of numerous global models suggests that the Toba Batak language has significant promise for preservation through the strategic integration of AI technology (Gulyamova and Rasulmuhamedova, 2025).

However, these implementations are largely supported by extensive linguistic resources and institutional funding, conditions that are not fully available for Toba Batak. Therefore, direct replication is not feasible, and adaptation strategies must consider data scarcity, community involvement, and morphological complexity.

3.3. Machine translation (MT) in Toba Batak language preservation

Machine translation (MT) for Toba Batak presents a fundamentally different set of challenges compared to high-resource languages due to its high morphological productivity and limited parallel corpora. Standard neural machine translation systems, which rely heavily on large-scale aligned datasets, are insufficient in this context because they treat words as atomic units rather than decomposable morphological structures. In Toba Batak, meaning is not encoded solely at the lexical level but distributed across complex affixal systems, including causative (*pa-...-hon*), passive (*i-*), and iterative forms (*-i*, reduplication). As a result, direct sequence-to-sequence translation models often fail to preserve semantic distinctions, particularly in voice alternations and argument structure, leading to inaccurate or flattened translations.

This limitation necessitates a shift toward morphology-aware MT architectures, where input sequences are enriched with morphological segmentation, tagging, or subword representations that reflect the internal structure of words. In low-resource scenarios, such approaches are not optional but essential for ensuring that translation outputs maintain both semantic fidelity and grammatical integrity.

AI-powered MT mechanically translates text between languages using statistical models, neural networks, and large linguistic datasets. Modern neural machine translation (MT) systems evaluate syntactic, semantic, and contextual data to produce translations that are about as accurate as human translations. Unlike traditional dictionary-based approaches, neural machine translation (MT) employs deep learning algorithms that can model complex speech patterns, making it an essential tool for multilingual digital communication. MT performance for low-resource languages, like minority and endangered languages, has improved recently. MT serves as a technological link between traditional linguistic knowledge and modern digital platforms in the context of language preservation (Broughton et al., 2015). MT enables the integration of languages with minimal documentation into global information networks by automating translation tasks and facilitating the creation of bilingual corpora, digital instructional resources, and parallel datasets. These findings support linguistic research, intergenerational communication, and community-driven revitalization initiatives.

The Austronesian language Toba Batak presents particular challenges for MT because of its complex affixation, agglutinative morphology, and voice alternations. Standard MT models sometimes struggle to understand such structures due to high morphological output and a paucity of training data. By combining morphology-aware neural architectures with community-contributed data, MT systems may handle canonical and derived forms more effectively, ensuring that Toba Batak's linguistic structural richness is preserved in digital contexts (Chen et al., 2023). MT aids in the creation of multilingual digital corpora that capture the essential linguistic features of Toba Batak. Automatically aligned source and target texts facilitate the development of parallel corpora for linguistic research, educational platforms, and cultural preservation initiatives. These corpora serve as the basis for other computational tools such as voice-activated applications, automatic glossing systems, and morphological analyzers, strengthening the digital infrastructure required for long-term language preservation.

Bilingual dictionaries, e-learning platforms, and cross-lingual chatbots are examples of MT applications that enhance accessibility for both native speakers and language learners (Alam, 2021). Translation of words, phrases, and culturally significant expressions in real time promotes daily usage and improves language skills. By making linguistic information easier to obtain and providing quick responses, MT encourages revitalization and community involvement in places where the number of fluent speakers is falling. Incorporating MT into Toba Batak language preservation projects fosters sustainable revival by encouraging documentation, instruction, and technical innovation. Through participatory data collection, where local speakers give recordings, translations, and morphological comments, MT systems can represent actual usage and cultural knowledge. By incorporating Toba Batak into digital infrastructures, boosting its visibility, and ensuring that linguistic knowledge is accessible to future generations, MT ultimately strengthens the language's resilience. Morphological examples in the Toba Batak language for MT operation are provided in Table 2 using the root word *allang* "eat."

The suffix *-hon* in *allanghon* above in the Toba Batak language functions as a derivational marker that forms new verbal stems. When attached to verb bases, this suffix undergoes morphophonemic alternation, yielding the allomorphs (1) –*kon*, (2) –

son, (3) –*ton*, and (4) –*pon*. The distributional rules governing these allomorphs can be formulated as follows:

- [-hon] → [-kon] when the final phoneme of the base is [k] or [g], as follows: -hon + *gok* "many" *gokhon* [gokkon] "to invite" -hon + *lonong* "to sink" → *lononghon* [lonokkon] "to submerge"
- [-hon] → [-son] when the root word ends in [s]: -hon + *rambas* "to slash" → *rambashon* [rabbasson] "to cause to slash" -hon + *hohos* "to tight" → *hohoshon* [hohosson] "to tighten"
- [-hon] → [-ton] when the root word ends in [t] or [n]: -hon + *surat* "to write" → *surathon* [suratton] "to write down" -hon + *suan* "to plant" → *suanhon* [suatton] "to plant"
- [-hon] → [-pon] when the root word ends in [m] or [p]: -hon + *gomgom* "to embrace" → *gomgomhon* [gomgoppon] "to cause to embrace" -hon + *golom* "to grasp" → *golomhon* [goloppon] "to cause to grasp"

The suffix –*an* in the Toba Batak language functions as a derivational marker for forming new lexical items. The attachment of –*an* to a root word exhibits a distinct pattern compared to other affixation processes. This distinction lies in its morphophonological behavior: the suffix –*an* does not produce allomorphic variants. Instead, it systematically triggers a phonemic alternation in which the final consonant [k] of the root word shifts to [h], as illustrated in Table 3.

Table 2: Morphological pipeline for machine translation (MT) based on the root word *allang* "to eat" in Toba Batak

Morphological construction (affixation process)	Surface form	Functional meaning
allang + -hon	allanghon	To feed someone (causative/imperative)
ma- + allang	mangallang	To eat (transitive verb)
ma- + allang + -i	mangallangi	To eat repeatedly (iterative transitive)
i- + allang	iallang	Eaten (passive/intransitive)
allang + -i	allangi	To eat repeatedly (imperative/iterative)
allang + -on	allangon	To be eaten (passive form)
si- + allang + -on	siallangon	Food (deverbal noun)
tar- + allang	tarallang	Accidentally eaten/can be eaten (stative/passive)
di- + allang + -hon	diallanghon	Eaten (passive causative)
allang-allang	allang-allang	To eat repeatedly (full reduplication)
di- + allang-allang	diallang-allang	To eat repeatedly (prefixed reduplication)
di- + allang-allang + -i	diallang-allangi	To eat repeatedly (circumfixed reduplication)

Table 3: Morphophonemic alternation of final /k/ shifts to /h/ triggered by the suffix *-an* in Toba Batak

Root word	Phonetic form	Lexical meaning	Suffix	Derived form (surface form)	Phonetic form	Functional meaning
hodok	[hodok]	sweat	-an	hodohan	[hodohan]	To perspire
ojak	[ojak]	stable	-an	ojahan	[ojahan]	Foundation
arsik	[arsik]	to drain	-an	arsihan	[arsihan]	To drain repeatedly
pultak	[pultak]	to break	-an	pultahan	[pultahan]	To break repeatedly

An effective MT system for Toba Batak language must be capable of interpreting a wide range of affixes (e.g., *ma-*, *ma-...-i*, *i-*, *on*, *si-...-on*, *pa-...-on*, *-hon*, *tar-*, *pa-*, *di-...-hon*, *di-...-i*), identifying reduplication patterns and maintaining semantic differences across formulations that are patient-focused, causal, and intransitive. In order to preserve the structural and semantic integrity of the Toba Batak language, the system must also provide syntactically accurate English translations that represent its morphological markers. These features are necessary to fully capture the intricacy of Toba Batak language expressions, as well as to facilitate the development of trustworthy bilingual corpora and AI-powered educational applications.

MT between Welsh and English has supported efforts to preserve and promote the Welsh language, which is another concrete example of AI-based translation technologies. In collaboration with academic institutions and the technology sector, the Welsh government has developed machine-learning translation platforms for media, education, and public services. These systems leverage extensive Welsh corpora and advanced computer models to ensure translation accuracy and contextual appropriateness. Nevertheless, these systems rely on large-scale annotated corpora and sustained institutional support, which remain limited in the Toba Batak context.

3.4 Text-to-speech (TTS) in Toba Batak language preservation

Text-to-Speech (TTS) systems for morphologically complex languages such as Toba Batak must address challenges that extend beyond phoneme-to-sound mapping. Unlike isolating languages, Toba Batak exhibits extensive morphophonemic alternations, where affixation not only modifies meaning but also systematically alters phonological realization. For example, suffix variation (e.g., *-hon → -pon, -ton, -son*) depends on the phonological properties of the root, requiring context-sensitive pronunciation modeling.

Conventional TTS architectures, particularly those trained on limited datasets, tend to produce phonologically inaccurate or morphologically flattened outputs, failing to reflect these alternations. This is especially problematic in language preservation contexts, where accurate reproduction of prosody, affix boundaries, and morphologically conditioned phonetic variation is essential.

Therefore, TTS development for Toba Batak should incorporate morphological parsing as a preprocessing stage, enabling the system to generate pronunciation not merely from orthography but from linguistically informed representations. This transforms TTS from a purely acoustic task into a morphology-phonology interface problem, which is critical for preserving both intelligibility and linguistic authenticity.

TTS technology, an essential component of AI, converts written text into spoken audio, offering critical applications for the preservation of endangered languages (Chen et al., 2023; Broughton et al., 2015). In this sense, TTS facilitates accurate replication of native pronunciation, improves language learning, and aids in the documentation of linguistic resources like dictionaries, morphological databases, and narratives. By creating audio output from written sources, TTS ensures that phonological components of the language are preserved and accessible. Additionally, it promotes regular use through interactive tools like educational platforms, reading programs, and chatbots. By recreating phonological traits that could otherwise be lost due to declining speaker populations, TTS strategically supports AI-driven language preservation by converting orthographic input into speech that sounds natural. TTS provides a sustainable means of transferring spoken language knowledge to future generations within preservation frameworks, enhances multimodal learning environments, and facilitates the creation of digital audio archives. Beyond its practical use, TTS is a crucial tool for maintaining a language's lexical, prosodic, and phonetic elements in digital settings (Chen et al., 2023). TTS is particularly helpful for morphologically complex languages such as Toba Batak language, which exhibit significant affixation, derivation, and morphophonemic alternations that produce several surface forms from a single lexical root. With AI-based TTS systems that employ morphological rules, learners can hear how prefixes, infixes, suffixes, and reduplication affect both pronunciation and semantic nuance. This ability is essential for revitalization efforts since many complex forms are underrepresented in everyday speech but are nevertheless essential to grammatical proficiency and cultural expression. When combined with morphology-based digital corpora, TTS enables interactive display of linguistic data in both textual and audio formats (Broughton et al., 2015). This connection enhances educational technologies such as digital dictionaries, chatbots, and literacy programs by providing real-time pronunciation for morphologically generated word forms. By transforming descriptive documentation into practical, interactive learning possibilities, this synergy makes morphological knowledge actively available to communities in minority language situations.

For every morphological diversity, TTS systems are able to produce accurate phonological outputs that demonstrate how affixation modifies meaning and auditory realization. By adding affixes to lexical roots, the agglutinative Toba Batak language creates new semantic forms. For example, TTS can examine the root word *torus*, which means "continue," to demonstrate how affixation impacts meaning and pronunciation, providing users with immediate auditory feedback and enhancing their understanding of morphological structures, as shown in Table 4.

Table 4: Morphology-aware text-to-speech processing based on the root word *torus* “to continue” in Toba Batak language

Morphological construction (affixation pattern)	Functional meaning	TTS processing description
i- + torus + -hon	To be continued (passive/progressive)	The TTS system generates a unified phonological output by processing the circumfix *i-...-hon* as a single morphological unit.
pa- + torus + -on	To cause something to continue (causative)	The TTS system models the causative circumfix *pa-...-on*, ensuring accurate pronunciation and prosodic integration with the root.
torus + -hon	To continue something (verbal derivation)	The TTS system distinguishes the suffix *-hon* and adjusts phonological realization accordingly.
tar- + torus + -hon	Can be continued (stative/passive)	The TTS system processes the circumfix *tar-...-hon* to generate morphologically and phonologically coherent speech output.

Another example, the formation of the intransitive verb *pa-tangkup-hon* [patakkuppon] in the Toba Batak language through the circumfix *pa-...-hon* proceeds through the following steps: circumfix *pa-...-hon* + *tangkup* → *patangkuphon* “arrest.” This configuration can be described as follows. The intransitive verb *patangkuphon* is derived from the root word *tangkup* “catch/arrest” combined with the circumfix *pa-...-hon*. Thus, the verb consists of two morphemes: the free morpheme *tangkup* and the bound morpheme *pa-...-hon*. The two bound elements function as an inseparable unit because they jointly encode the grammatical and lexical meaning of the construction. The intransitive verb *patangkuphon* is derived from the free root word *tangkup*. Its grammatical meaning denotes “causing another person to perform the action expressed by the root word.” The lexical meaning of *patangkuphon* is “to have someone capture (someone else),” indicating the involvement of an agent who is instructed or hired to carry out the capture. The segment –*hon* in the circumfix *pa-...-hon* appears as –*pon* because the final phoneme of the root word *tangkup* is [p]. The same morphophonemic alternation occurs with root words ending in [m], where phoneme [h] alters to [p] as illustrated in Table 5. A TTS system can convert written input into natural spoken Toba Batak language, preserving both intonation and the morphological structures embedded in root words, such as *torus*, along with affixes including *i-...-hon*, *pa-...-on*, *-hon*, and *tar-...-hon.* As an endangered language with active preservation efforts, the Toba Batak language encodes both cultural and functional meaning within its morphological system. By generating all possible morphological variants, including forms rarely used in daily speech, TTS facilitates learners' understanding of productive affixation, maintains the phonetic integrity of each morpheme, and supports the development of morphology-based digital corpora and AI-driven educational applications.

Table 5: Morphophonemic alternation of the suffix *-hon* shifts to */-pon/* in circumfix *pa-...-hon* constructions

Morphological construction	Surface form	Phonetic form	Functional meaning
pa-...-hon + gomgom	pagomgomhon	[pagomgoppon]	To cause someone to embrace
pa-...-hon + golom	pagolomhon	[pagoloppon]	To cause someone to grasp
pa-...-hon + golap	pagolaphon	[pagolappon]	To darken/to cause something to become dark

3.5. Chatbots in Toba Batak language preservation

The development of AI-driven chatbots for Toba Batak requires careful consideration of the language's morphosyntactic complexity and low-resource constraints. While modern conversational systems are typically powered by large language models, their effectiveness is significantly reduced when applied to languages with limited training data and highly productive morphology.

In Toba Batak, conversational meaning is often encoded through derivational processes, voice marking, and reduplication, which influence not only lexical meaning but also pragmatic interpretation. Standard chatbot architectures, which rely on surface-level pattern matching or pre-trained embeddings, are insufficient for capturing these nuances, resulting in responses that may be grammatically incorrect or semantically incomplete.

This limitation highlights the need for morphology-aware conversational models, where user input is first analyzed into its morphological components before semantic processing. Such an approach enables the system to generate responses that are structurally accurate and contextually appropriate, while also contributing to the incremental construction of annotated linguistic corpora through user interaction.

Chatbots are AI-powered conversational devices that simulate human interaction through machine learning, natural language processing, and knowledge-based reasoning (Chen et al., 2023). Modern neural chatbots employ deep learning architectures that dynamically adapt to user input, in contrast to traditional rule-based systems. This enables contextually relevant responses and interactive task performance across digital platforms. These characteristics make chatbots particularly useful for multilingual communication, educational applications, and digital information retrieval. In the realm of language preservation, chatbots are readily accessible digital tools that promote active engagement with minority and endangered languages. By offering lexical support, real-time conversational practice, and culturally embedded phrases, chatbots enable younger speakers to interact with their heritage language in pertinent situations. This interactive element encourages intergenerational transmission and cultural continuity by transforming static documents into dynamic language experiences.

Morphology-aware chatbots offer a novel way to record and share linguistic information for the Toba Batak language, an Austronesian language with agglutinative morphology, voice alternations, and productive affixation. By simulating intricate verbal and nominal structures like derivation, reduplication, and focus marking, these systems help preserve structural depth that could otherwise be lost as a result of younger generations' deteriorating fluency or a lack of formal educational resources. By gathering conversational data during encounters, chatbots also aid in the creation of language corpora. By annotating these datasets for phonology, morphology, syntax, and semantics, digital documentation can be improved, and complementary AI systems like speech recognition, text-to-speech, and MT customized for the Toba Batak language can be trained. This feedback loop enables chatbots to both utilize and expand linguistic resources.

Chatbots in educational applications offer interactive learning settings with real-time feedback, pronunciation assistance, and morphologically correct examples. In addition to supporting everyday language use in communities undergoing a shift toward Indonesian or English, they serve as digital tutors for vocabulary development, sentence building, and culturally grounded phrases. A sustainable and inclusive approach to digital revival is provided by incorporating chatbot technology into Toba Batak language preservation initiatives. Chatbots provide chances for participatory data gathering, supplement current documentation, and enhance language practice by interacting with speakers in genuine conversation. The digital preservation of Toba Batak language displays genuine, community-driven linguistic knowledge thanks to community contributions of vocabulary, expressions, and culturally meaningful narratives. Chatbot practices in Toba Batak language are shown in Table 6 using the root word *gelleng* "small."

Furthermore, the process of word formation that yields a new lexical category is derivational in nature. For instance, the verb *hatuhil* [hatuhil] is formed from the root word *tuhil* "chisel" combined with the bound morpheme ha-. The grammatical meaning encoded in this formation expresses "an action that can be performed using the instrument denoted by the root word."

Consequently, the lexical meaning of the derived verb *hatuhil* [hatuhil] is "to chisel." Other examples are presented in Table 7.

Table 6: Morphological patterns for chatbot processing using the root *gelleng* "small" in Toba Batak language

Morphological process	Surface form	Functional meaning
um- + gelleng	Unggelleng	Smaller
gelleng + -an	Gellengan	Smaller
um- + gelleng + -an	Unggellengan	Comparatively smaller
pa- + gelleng	Pagelleng	To make smaller
dipa- + gelleng	Dipagelleng	To be made smaller
pina- + gelleng	Pinagelleng	Has been made smaller
pa- + gelleng + -hon	Pagellenghon	To cause to become smaller
gelleng-gelleng	Gelleng-gelleng	Repeatedly small/distributed smallness
pa- + gelleng-gelleng	Pagelleng-gelleng	To gradually reduce size

Table 7: Derivational verb formation with the prefix *ha-* from nominal roots in Toba Batak language

Root word	Derived form	Grammatical function	Lexical meaning	Category shift
Labang	Halabang	Instrumental derivation	To be nailed/can be nailed	Noun → verb
Hihir	Hahihir	Instrumental derivation	To be filed/can be filed	Noun → verb
Tongkok	Hatongkok	Instrumental derivation	To be hammered/can be hammered	Noun → verb

An effective Toba Batak language chatbot must be capable of recognizing a wide range of affixal forms (e.g., *um-*, *-an*, *um-...-an*, *pa-*, *dipa-*, *pina-*, *pa-...-hon*), interpreting reduplication patterns (e.g., *gelleng-gelleng*), distinguishing between morphological voice types (agentive, passive, causative), and generating contextually appropriate English equivalents. These functionalities are essential to ensure that chatbot-mediated interactions accurately reflect both the semantic nuances and the morphological complexity of Toba Batak language, thereby supporting robust language learning, digital documentation, and community-driven preservation initiatives. AI has been applied in New Zealand to enhance the learning and preservation of Māori, an Indigenous language that is vital to the country's cultural identity (Moshagen et al., 2024). One well-known effort is the development of AI-based chatbots specifically designed to assist Māori language learners through conversational exchanges (Broughton et al., 2015). By combining Māori cultural and historical knowledge with language exercises, these AI systems offer a more thorough and culturally important learning experience. This digital strategy has successfully attracted younger students and internet users by offering a more accessible and interesting alternative to traditional teaching methods (Alam, 2021). While effective, such implementations require substantial data and infrastructure, highlighting the need for adapted, resource-aware approaches for Toba Batak. The success of New Zealand's AI-supported language education initiative shows how important it is to combine technical innovation with culturally grounded material to increase participation and relevance in minority language teaching.

4. Conclusion

This study proposes a morphology-aware computational framework for Toba Batak language preservation, supported by a pilot implementation that demonstrates its practical feasibility. By integrating linguistic knowledge with AI architectures, the framework addresses key challenges associated with agglutinative, low-resource languages, including data sparsity and morphophonemic complexity. The results highlight the importance of embedding morphological intelligence into AI systems to ensure linguistic accuracy and sustainability. This approach provides a scalable model for other under-resourced languages, contributing to broader efforts in AI-driven language preservation.

List of abbreviations

AI	Artificial intelligence
ASR	Automatic speech recognition
BPE	Byte pair encoding
FST	Finite-state transducer
FSTs	Finite-state transducers
MT	Machine translation
NLG	Natural language generation
NLP	Natural language processing
POS	Part-of-speech
TTS	Text-to-speech

Funding

This research was funded by the Direktorat Penelitian dan Pengabdian kepada Masyarakat, Direktorat Jenderal Riset dan Pengembangan, Kementerian Pendidikan Tinggi, Sains, dan Teknologi Republik Indonesia, under Master Contract No. 122/C3/DT.05.00/PL/2025 and Derivative Contract No. 5/SPK/LLI/AL.04.03/PL/2025.

Acknowledgment

The authors would like to express their sincere gratitude to the Direktorat Penelitian dan Pengabdian kepada Masyarakat, Direktorat Jenderal Riset dan Pengembangan, Kementerian Pendidikan Tinggi, Sains, dan Teknologi Republik Indonesia, as well as the Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) Universitas Methodist Indonesia, for their valuable support in this research project.

Compliance with ethical standards

Ethical considerations

This study involved human participants. Informed consent was obtained from all informants prior to data collection. The confidentiality and anonymity of participants were strictly maintained.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Alam A (2021). Possibilities and apprehensions in the landscape of artificial intelligence in education. In the 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), IEEE, Nagpur, India: 1-8. https://doi.org/10.1109/ICCICA52458.2021.9697272

Ambarita E and Sembiring M (2025). The patterns of ablaut reduplication in Toba Batak: A construction-based approach. REiLA: Journal of Research and Innovation in Language, 7(1): 29-44. https://doi.org/10.31849/reila.v7i1.17024

Broughton D, Te Aitanga-a-Hauiti T, Porou N, McBreen K, Waitaha KM, and Tahu N (2015). Mātauranga Māori, tino rangatiratanga and the future of New Zealand science. Journal of the Royal Society of New Zealand, 45(2): 83-88. https://doi.org/10.1080/03036758.2015.1011171

Chen Y, Jensen S, Albert LJ, Gupta S, and Lee T (2023). Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. Information Systems Frontiers, 25: 161-182. https://doi.org/10.1007/s10796-022-10291-4

Gulyamova G and Rasulmuhamedova D (2025). Integration of artificial intelligence into educational platforms for effective language learning. Jurnal Ilmiah Ilmu Terapan Universitas Jambi, 9(3): 1009-1024. https://doi.org/10.22437/jiituj.v9i3.42134

Kim J, Merrill Jr K, and Collins C (2021). AI as a friend or assistant: The mediating role of perceived usefulness in social AI vs. functional AI. Telematics and Informatics, 64: 101694. https://doi.org/10.1016/j.tele.2021.101694

Liu GL, Darvin R, and Ma C (2025). Exploring AI-mediated informal digital learning of English (AI-IDLE): A mixed-method investigation of Chinese EFL learners' AI adoption and experiences. Computer Assisted Language Learning, 38(7): 1632-1660. https://doi.org/10.1080/09588221.2024.2310288

Low DS, McNeill I, and Day MJ (2022). Endangered languages: A sociocognitive approach to language death, identity loss, and preservation in the age of artificial intelligence. Sustainable Multilingualism, 21(1): 1-25. https://doi.org/10.2478/sm-2022-0011

Moseley C (2010). Atlas of the world's languages in danger. UNESCO, Paris, France.

Moshagen SN, Antonsen L, Wiechetek L, and Trosterud T (2024). Indigenous language technology in the age of machine learning. Acta Borealia, 41(2): 102-116. https://doi.org/10.1080/08003831.2024.2410124

Salazar Alva RG, Núñez M, López del Alamo C, and Martín Arista J (2026). AI-driven generation of old English: A framework for low-resource languages. Big Data and Cognitive Computing, 10(5): 145. https://doi.org/10.3390/bdcc10050145

Sumandya IW, Widana IW, Hendra R, Supian S, Yusuf M, Wijaya HA, and Safitri RA (2025). Adoption of artificial intelligence in STEM learning: Examining the effects of performance, effort, and social factors. Jurnal Ilmiah Ilmu Terapan Universitas Jambi, 9(4): 1341-1354. https://doi.org/10.22437/jiituj.v9i4.44475