Sinhala Speech to Sinhala Sign Language Translation System Using Human Interpreter Video Synthesis for Inclusive Communication

Interim Report

SCS 4224: Final Year Project in Computer Science

M.K.P. Ahinsa

2021/CS/003

University of Colombo School of Computing

Signatures of Researcher and Supervisors

Ms. M.K.P. Ahinsa
pathuahinsa2001@gmail.com
Researcher

Dr. Kasun Karunanayaka
ktk@ucsc.cmb.ac.lk
Supervisor

Mrs. Sanduni Thrimahavithana
sst@ucsc.cmb.ac.lk
Co-Supervisor

Abstract

This is the interim report for the research titled "Sinhala Speech to Sinhala Sign Language Translation System Using Human Interpreter Video Synthesis for Inclusive Communication." The report presents the motivation and background for addressing communication barriers faced by the Sri Lankan deaf community, emphasizing the limited accessibility of Sri Lankan Sign Language (SSL) translation tools. It outlines the central research problem, identifies gaps in existing work, and defines the scope and delimitations of the project. The proposed solution involves developing a system that translates Sinhala text and speech into SSL animations using human interpreter video synthesis, leveraging techniques from natural language processing, machine translation, and computer vision. The report also details the research methodology, including data collection, gloss generation, parallel corpus creation, and baseline model development, and presents preliminary results. Finally, it describes the planned evaluation strategies to assess system effectiveness and usability, ensuring the research contributes to inclusive communication and social integration for the deaf community.

Contents

List of Figures

List of Tables

List of Abbreviations

ASL American Sign Language.

ASR Automatic Speech Recognition.

BP Brazilian Portuguese.

BSL British Sign Language.

CBMT Corpus-based Machine Translation.

CNNs Convolutional Neural Networks.

DHH Deaf and Hard of Hearing.

DSR Design Science Research.

HamNoSys Hamburg Notation System.

NLP Natural language processing.

NLTK Natural Language Toolkit.

NMT Neural Machine Translation.

RBMT Rule-based Machine Translation.

RNNs Recurrent Neural Networks.

SSL Sri Lankan Sign Language.

STT Speech to Text.

1 Introduction

1.1 Background

The deaf community uses hand gestures and facial expressions to convey their ideas to people with hearing impairments or those with normal hearing. That language which uses hand gestures and facial expressions to communicate is called sign language, and this is not an international language of hearing-impaired individuals. Different countries have their sign languages, often with various dialects within the same country (Groundviews n.d.). These boundaries do not necessarily align with spoken languages or national borders. There are currently nearly a hundred sign languages recognized worldwide; among them, American Sign Language (ASL) and British Sign Language (BSL) are two of the most structured and well-defined. Most spoken languages have sign languages that have evolved or been developed based on established ones like ASL and BSL (Fernando & Wimalaratne 2016).

The Sri Lankan deaf community uses the Sri Lankan Sign Language (SSL) as their basic language. SSL was originally influenced by BSL because the very first deaf school was established during the colonial period. However, several distinct sign languages have developed across different regions, mainly because many hearing-impaired individuals, especially in rural areas, lack access to formal education. This leads to the use of informal and regionally varied sign gestures (Daily Mirror 2018).

Globally, the field of sign language translation has witnessed progress through the use of Natural Language Processing (NLP), computer vision, speech recognition, and 3D avatar animation. Research efforts have produced real-time translation systems for ASL and BSL using deep learning, convolutional neural networks (CNNs), recurrent neural networks (RNNs), and rule-based grammar parsers.

In Sri Lanka, research on SSL remains sparse. Early studies have mostly focused on recognizing SSL signs using computer vision and wearable sensors (Sewwantha & Ginige 2021) (Munasinghe et al. 2023). Few have attempted end-to-end systems for translating Sinhala text or speech into SSL. Additionally, most of the local research efforts are confined to limited domains (e.g., education, medical terms) and use constrained vocabularies, lacking the scalability required for practical use.

1.2 Challenges

In the Sri Lankan context, there are over 400,000 deaf and hearing impaired individuals (Daily Mirror 2018). SSL was officially recognized in September 2017 through the Conversational Sign Language Bill, making Sri Lanka one of only 39 countries worldwide to officially recognize a national sign language.

Generally, it is difficult for normal-hearing people to understand this sign language without proper training. As a result, there is a communication gap between these hearing-impaired and normal-hearing people, especially related to public services such as banks, hospitals, police stations, etc. Also, Sri Lanka faces a severe shortage of SSL interpreters, with only six practising interpreters

as of 2018. As a result, those who use sign language often experience social isolation and may feel overwhelmed in their daily lives (Daily Mirror 2018).

Another significant challenge for deaf individuals is their inability to communicate with others who are geographically distant. In contrast, normally hearing individuals have various means of interaction through social networking and chat applications. Although tools like Skype facilitate communication among deaf or mute individuals who use sign language, they do not support effective communication between normally hearing and deaf persons due to the lack of translation facilities (Fernando & Wimalaratne 2016).

1.3 Motivation

The marginalization of the deaf community in Sri Lanka stems from both social barriers and technological gaps. While global research is being conducted to bridge communication gaps in various sign languages, low-resource sign languages like SSL do not have sufficient solutions suggested for this area. Among them, most of the exclusion restricts deaf individuals' access to essential services and their full participation in society.

An automated system that translates Sinhala text and speech into SSL animations could significantly enhance inclusivity. It would allow hearing individuals to communicate effectively with the deaf community, reducing dependence on human interpreters and improving access to services, education, and information for deaf persons.

This situation underscores the urgent need for technology-driven solutions that can bridge this communication divide. Like other sign languages, SSL relies on a unique combination of hand gestures, body movements, and facial expressions for communication. Consequently, ideas expressed or messages conveyed by a person with a hearing or speech impairment using SSL are often difficult for those without knowledge of SSL to understand. Since most normally hearing (NH) individuals do not understand sign language, deaf persons frequently find it difficult to navigate everyday tasks, particularly when engaging with public services such as banks, hospitals, and police stations. This creates a significant communication barrier between deaf individuals and those who can hear, unless an interpreter is available.

2 Scope, Delimitation and Justification

2.1 Scope

This project focuses on developing an end-to-end Sinhala (audio/text) to SSL translation system tailored to the communication needs of the Sri Lankan Deaf community. The system aims to generate accurate, understandable, and culturally appropriate SSL gestures using stitched human interpreter videos. The scope of the project includes the following components:

- Input pre-processing
  - Speech-to-Text Conversion: Accurately convert spoken Sinhala audio to text using

noise-robust, accent-tolerant speech recognition models (e.g., enhanced Web Speech API or deep learning-based ASR models).

  - Text Preprocessing: Normalize Sinhala text inputs to handle punctuation, word segmentation, tokenization, and morphological decomposition (handling tenses, case markers, verb conjugations, etc.).

- Neural Machine Translation

  - Parallel Corpus Creation: Construct and align a Sinhala-to-SSL gloss corpus, including core vocabulary(nouns/verbs/adjectives) based on the Sri Lankan Sign Dictionary and expert-reviewed translations.
  - Model Selection and Evaluation: Experiment with and evaluate multiple NMT architectures (e.g., Transformer(mBART,mBert) to determine the most effective model for Sinhala → SSL gloss translation based on BLEU score, gloss sequence accuracy, and semantic preservation.
  - Handling Linguistic Complexity: Ensure the NMT model handles agglutinative Sinhala grammar, multi-word expressions, and context-sensitive meanings with techniques such as subword tokenization.

- Incorporate fingerspelling for the Sinhala alphabet and numerical signs, following the dictionary's guidelines.

- Generate outputs as stitched human interpreter videos, validated against the dictionary's standards.

2.2 Delimitations

The project explicitly excludes tasks and features beyond its core objective. Reverse translation (SSL to spoken/written Sinhala) and real-time sign recognition (interpreting SSL from video inputs) are not included. The system does not support translation to or from other sign languages (e.g., Tamil Sign Language) or spoken languages outside Sinhala.

2.3 Justifications

The Sri Lankan Sign Dictionary serves as the authoritative reference for this project, ensuring linguistic consistency and reducing ambiguity in dataset creation and model training. By prioritizing spoken-to-sign translation, the system directly addresses the Deaf community's urgent need to access spoken information independently. This focused approach also establishes a standardized foundation for future expansions, such as incorporating additional translation directions, without compromising the initial deliverables.

3 Preliminary Literature Review

Spoken-to-sign language translation has emerged as a critical area of research worldwide, driven by the need to bridge communication gaps for deaf communities. Although advances in technologies such as NMT and generative adversarial networks (GANs) have revolutionized sign language production for languages such as ASL and BSL, significant challenges persist in low-resource contexts like SSL. This review synthesizes global methodologies and Sri Lankan advancements, highlighting gaps that motivate the proposed research. A comprehensive review of spoken-to-sign language translation systems was conducted and published by me in the journal FAITH (Ahinsa et al. 2025), providing a detailed analysis of the evolution, challenges, and future directions of this field. In addition, a dedicated review of SSL technologies and resources was prepared[1]. These works serve as the foundation for the present research and are referenced throughout this proposal.

According to our literature survey and other related reviews, most sign language translation systems use three main steps to translate spoken language into sign language: text/audio preprocessing, machine translation, and sign synthesis.

3.1 Text/Audio Preprocessing

Preprocessing involves two major tasks: converting speech to text and preparing that text for translation. Automatic Speech Recognition (ASR) tools, such as Google's Speech-to-Text API, are commonly used to convert spoken input into textual data. Patel et al. (2025) explored the conversion of speech and text into American Sign Language (ASL) and Indian Sign Language. The user input can be provided either through speech or by typing directly into a designated input field. When speech input is used, the Web Speech API transcribes it in real time, converting the audio into text that can be processed further. However, this audio-to-text translation faces challenges due to background noise cancellation issues. The researchers suggested that enhancing advanced noise cancellation techniques with AI could improve the accuracy of the transcription.

In the text/audio preprocessing stage, systems use various NLP tools such as NLTK, Stanford Parser, MWETokenizer, CoreNLP, and WordNet to provide automated text processing solutions. Some systems have extended these NLP preprocessing techniques to multiple languages (e.g., Hindi, Telugu, and English), making the conversion more comprehensive (Rao et al. 2023). Although tokenization and stop-word removal methods improve conversion efficiency, they sometimes change or oversimplify the meaning of input sentences. Some systems manually handle stop-word removal to enhance performance. However, many tokenization tools are limited to English and do not accurately tokenize rural languages like Sinhala, which has a complex grammatical structure. Some research has described the limitations of SSL fingerspelling when tokenizing, noting the lack of single fingerpelling signs to animate conjuncts, which are not directly listed in the Sinhala alphabet. Modifiers like repaya, which modify consonant sounds or reduce syllables, are also not visually represented in gesture animations. In the system presented by (Punchimudiyanse

[1] https://drive.google.com/file/d/1LPYGENzfXnPzR85wjlcZtqMgSTVm9w1P/view?usp=sharing

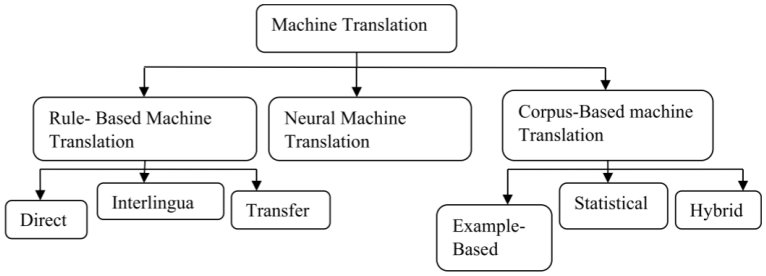

Figure 1: Taxonomy of machine translation methods (Kahlon & Singh 2023)

& Meegama 2017*a*), only 61 SSL fingerspelling signs were supported, relying on phonetic English conversion and pre-rendered 3D animations. This rule-based, phonetic approach fails to capture key morphological and syntactic constructs of Sinhala—particularly conjunct letters and modifier rules—leading to unnatural or incomplete representations. Therefore, text processing has many limitations for low-resource sign languages like Sinhala, Indian, and Arabic (Punchimudiyanse & Meegama 2017*a*).

3.2 Sign Sequence Generation Using Machine Translation Methods

In the machine translation step, three main types of methods can be identified: rule-based, corpus-based, and neural machine translation methods such as in Figure 1 (Ahinsa et al. 2025) (Alaghband et al. 2023).

Most previous research uses rule-based machine translation, which relies on well-defined grammatical and syntactic rules to handle structured sentences. Systems such as ATLASLang MTS (Brour & Benabbou 2019) and ASLG-PC12 (Tmar et al. 2013) use morpho-syntactic analysis to categorize words based on gender, number, tense, etc. These systems allow linguistic experts to modify or add new lemmas and linguistic rules, making them customizable and adaptable to evolving vocabularies. This flexibility is particularly beneficial for low-resource sign languages, such as Sinhala, Arabic, and Pakistani sign languages, which often evolve by creating new words. However, rule-based machine translation faces significant scalability challenges in development and maintenance, as more human effort and time are needed to create and validate rules. Managing infinite linguistic variations complicates the handling of complex sentences. For example, an artificial English-ASL corpus building system encountered difficulties validating all transformation rules, focusing only on essential parts, such as negative sentences and WH-word management (Tmar et al. 2013). (Ghosh & Mamidi n.d.) addressed this by using synonym substitution and multiword expressions, achieving more accurate results (95.833%) for unseen sentences than other systems. Although synonym substitution improves performance for Indian Sign Language, it may cause errors when synonyms do not convey the exact meaning of the source text. The ATLASLang project partially mitigates rule-based limitations by integrating an example-based fallback approach, but it still struggles with unseen data, limiting flexibility.

Corpus-based machine translation (CBMT) significantly improves over rule-based methods by using examples or statistical decision theories rather than hand-made rules. CBMT includes sta-

tistical, example-based, and hybrid approaches, each with unique limitations. Systems translating the English-to-Dutch sign language (Morrissey & Way 2009) and the French-to-Azee expression (Bertin-Lemée et al. 2023) have shown success in sentence structure retrieval and recombination using example-based methods. However, reliance on pre-existing corpora limits their handling of unseen words and structures and their ability to represent non-manual features like facial expressions, classifiers, and spatial relationships, which are crucial in sign language grammar but difficult to encode in text-based corpora. Thus, although CBMT provides a solid foundation, corpus limitations, lack of semantic depth, and sign language grammar complexity limit its effectiveness.

NMT is an emerging and promising approach to sign language translation systems. Many current research efforts use deep learning and transformer-based models, outperforming other data-driven or rule-based methods. For example, the Brazilian Portuguese (BP) to Libras translation project addresses data scarcity by pre-training on a high-resource language pair (Portuguese-English) before fine-tuning on BP-Libras with transformer models (Martino et al. 2023). However, this assumes spoken languages are easily adaptable to sign languages, which is not accurate since sign languages rely heavily on gestures, facial expressions, and spatial grammar.

A critical factor in NMT performance is the quality of contextual embeddings. For languages like English, large pretrained embeddings capture syntactic and semantic nuances effectively. In contrast, Sinhala suffers from a lack of high-quality, large-scale embeddings and annotated corpora, making it difficult to generate accurate contextual representations. Without robust embeddings, NMT models may fail to correctly encode meaning, particularly for morphologically rich constructs, conjunct letters, and modifiers unique to Sinhala. This limitation often propagates errors to the sign generation stage, reducing the naturalness and understandability of the output.

Another challenge is the use of glosses as intermediaries, which can lose semantic nuances and cause inaccuracies. Additionally, neural machine translation for sign language is still under development globally, and no such research has been reported for SSL yet. The recent work by Martino et al. (2023) demonstrates the feasibility of neural machine translation for text-to-sign using encoder-decoder architectures and attention mechanisms, achieving promising results for other languages, highlighting a gap and opportunity for SSL.

3.3 Sign Synthesis

The final stage of a text-to-sign translation system is sign synthesis, where generated gestures are visualized in a way that is understandable and expressive to the deaf community. Two main approaches dominate this area: avatar-based animation and video-based synthesis.

Early systems often relied on avatar-based methods, where 3D models were rigged and animated to perform sign gestures. For example, systems such as SignSynth (Webd & Grieve-Smith 2001) and the Sinhala Sign Language systems by Punchimudiyanse & Meegama (2017*b*),Punchimudiyanse & Meegama (2015) and Thrimahavithanaa et al. (2019) utilized platforms like Blender, MakeHuman, and Java3D to animate fingerspelling and basic signs. These systems involved mapping Sinhala Unicode to phonetic English and manually interpolating bone positions using Python scripts. While

they enabled structured rendering of signs and modifiers, such avatars typically lack natural fluidity, facial expressions, and fine hand articulation, making them less realistic than human-performed signing.

In contrast, more recent efforts have shifted toward video-based synthesis, aiming to produce outputs that are visually more similar to those of human interpreters. For instance, Stoll et al. (2018) introduced a GRU-based NMT model with Luong attention, generating gloss sequences from spoken or written input, followed by pose-conditioned GANs to synthesize HD skeletal videos. These GAN-based models use OpenPose joint coordinates and signer images to generate continuous, multi-signer sign language videos, offering significant improvements in naturalness and expressiveness compared to rigid avatars. However, challenges remain in capturing fine-grained hand movements and achieving photorealistic resolution, which are critical for fully intelligible signs.

Other systems, such as ATLASLang MTS1 (Brour & Benabbou 2019), begin with GIF-based representations and aim to evolve into more dynamic 3D animations, while Delorme et al. (2009) proposed a geometric description model with inverse kinematics for smoother transitions, yet still operate within the avatar framework. A hybrid system for Indian Sign Language by Gupta et al. (n.d.) combines pre-recorded video clips with synthetic animations based on HamNoSys notation, showing that many real-world applications still depend on concatenated video segments due to their higher realism compared to synthetic avatars.

This study does not focus on avatar generation but rather aligns with recent trends in natural video-based gesture synthesis, recognizing that human-like expressiveness, including subtle facial expressions, eye gaze, and body posture, remains underrepresented in most current systems. These non-manual components are crucial for conveying full semantic meaning and emotional tone in sign language, yet are often missing in both avatar and basic video concatenation approaches.

In response, AI-NLP-integrated systems offer promise in bridging this gap. With contextual language modelling, speech recognition integration, and multimodal input processing, such systems could pave the way for fluid, expressive, and human-friendly sign language synthesis. However, building such systems for low-resource sign languages like SSL requires addressing foundational challenges in data scarcity, multimodal annotation, and signer variability (Punchimudiyanse & Meegama 2017*b*).

4 Problem Statement and Research Gaps

4.1 Central Research Problem

The main problem is the communication gap between hearing people and the deaf community in Sri Lanka. Hearing people often struggle to express their ideas to deaf individuals because there are no proper tools to translate Sinhala into SSL. Existing systems are basic, do not include facial expressions, and cannot handle Sinhala's word changes like case markers and verb forms. This causes incorrect translations and limits access to important services for the deaf community.

4.2 Gaps in Existing Work

Despite advances in sign language translation systems globally, significant gaps remain, particularly in the context of SSL. These gaps limit the development of accurate and user-friendly translation systems for the deaf community in Sri Lanka.

- Scarcity of large annotated datasets for training and evaluating such models.
  - There is a lack of large, high-quality parallel datasets mapping Sinhala audio/text to SSL glosses and gestures.
  - Lack of enough good data makes it hard to train deep learning models that can translate languages accurately and reliably.
- Lack of methods for mapping Sinhala grammatical constructs to SSL equivalents.
  - Current systems are primarily rule-based, struggling with Sinhala's agglutinative morphology, including complex verb conjugations, case markers, and contextual syntax.
  - Existing rule-based systems do not generalize well to out-of-vocabulary terms or flexible sentence structures, leading to semantic loss and poor translation accuracy.
- Lack of NMT models tailored for low-resource languages such as Sinhala and SSL.
  - Currently, there are no NMT models specifically designed for the Sinhala-SSL pair.
  - Most neural systems rely on well-annotated parallel corpora, which are scarce or nonexistent for SSL.
- Limited incorporation of non-manual features (facial expressions, body posture) in avatar synthesis.
  - Sign language relies heavily on non-manual markers like facial expressions, lip movements, and body posture to convey meaning.
  - Current systems lack FACS-based animation or vision-based analysis to reproduce these critical elements.
- Low Realism in Existing Gesture Representations.
  - Most SSL systems use predefined GIFs or stitched videos, resulting in unnatural transitions and rigid expression.
  - There is no widely adopted 3D avatar system tailored for SSL that incorporates real-time, dynamic gesture synthesis.

4.3 Significance of the Research

This research holds significant importance in advancing inclusive communication technologies, particularly for the deaf community in Sri Lanka and other regions using low-resource sign languages like SSL. By addressing the critical gaps identified in current systems, this study aims to contribute to several key areas:

- By compiling and annotating comprehensive parallel corpora of Sinhala audio/text and SSL glosses, this project will create valuable datasets that can be leveraged by future researchers and developers, catalyzing further innovation in the field.
- Developing novel translation methods tailored to Sinhala's complex linguistic structure will enable more accurate and meaningful communication between Sinhala speakers and SSL users, fostering greater social inclusion and accessibility.
- This research will be the first to design and implement neural machine translation models specifically for the Sinhala–SSL pair, overcoming the challenges posed by limited annotated data through innovative approaches.
- Instead of relying on traditional avatar-based animation, this study explores the generation of sign gestures through natural human video outputs enriched with facial expressions and visual clarity. This approach is expected to produce more realistic, human-friendly translations, improving the comprehensibility and emotional nuance of the communication.
- The methodologies developed here can serve as a blueprint for other low-resource sign language pairs worldwide, promoting broader adoption of inclusive technologies in underrepresented communities.

Overall, this research promises to advance the state of spoken-to-sign language translation in Sri Lanka, empower the deaf community by facilitating seamless communication, and contribute novel scientific knowledge and practical tools to the global field of sign language technology.

5 Research Objectives and Questions

5.1 Aims

To develop a multimodal neural translation system that converts Sinhala audio/text into accurate, linguistically valid SSL gestures, validated through technical and user-centric evaluation.

5.2 Objectives

- To build a custom parallel dataset consisting of Sinhala sentences paired with corresponding SSL glosses.
  - Use expert translators and community contributions to annotate data.

  - Ensure coverage of formal, educational, and commonly spoken Sinhala phrases.

- To integrate speech-to-text transcription to enable voice-based input for real-time translation.
  - Analyze speech input in noisy and accented environments.
  - Apply preprocessing for better transcription accuracy.

- To design and implement a Sinhala-to-SSL word mapping model using NMT techniques.
  - Handle complex Sinhala grammatical constructs such as case markers, tense, and verb conjugations.
  - Explore subword tokenization strategies to manage agglutinative morphology.

- Generate and implement a human-friendly output using stitched human interpreter videos to provide natural and accurate Sinhala Sign Language translation, and evaluate this method for its accuracy, usability, and scalability in comparison to other approaches.

5.3 Research Questions

- **RQ 01:** How can a scalable Sinhala–SSL parallel gloss and video dataset be constructed to support expressive and accurate sign language translation?
  - Justification: Existing datasets are limited in scope, lack high-quality SSL video samples, and are not designed for Sinhala language translation. They do not adequately represent the vocabulary, grammar, or non-manual features necessary for accurate SSL synthesis.
  - Scope: This research focuses on building a domain-diverse, gloss-aligned SSL dataset using real human interpreters, covering a wide range of Sinhala vocabulary, including formal, informal, and commonly used expressions.

- **RQ 02:** Which techniques are most suitable for enhancing the accuracy of Sinhala speech-to-text (STT) systems?
  - Justification: Speech recognition for Sinhala is still maturing, and real-world scenarios often involve background noise, regional accents, and low-resource audio conditions. Investigating effective preprocessing (e.g., noise filtering, voice activity detection) is essential to ensure robust STT accuracy.
  - Scope: Focus on comparing preprocessing pipelines and their impact on downstream STT accuracy using real-world Sinhala voice samples.

- **RQ 03:** How can NMT be used to map Sinhala text inputs, including complex grammatical forms like case markers and verb conjugations, to accurate SSL gloss sequences?

  - Justification: Existing rule-based systems fail to handle Sinhala's agglutinative morphology and complex grammar, resulting in inaccurate or incomplete translations. Neural approaches, particularly transformers, offer context-aware modelling to preserve semantic modeling during translation.
  - Scope: Focus on developing and evaluating NMT models for Sinhala-to-SSL gloss translation, including subword tokenization and handling of grammatical variations (tenses, cases).

- **RQ 04:** How can sign language output be generated in a natural and easily understandable way using stitched human interpreter video segments?
  - Justification: Sign language users prefer human-interpreter visuals because they include important features like facial expressions, emotions, and smooth movements. Using pre-recorded video segments helps preserve this natural style, but stitching them together smoothly remains a challenge, especially for timing and flow.
  - Scope: This research focuses on evaluating the clarity, expressiveness, and fluency of SSL outputs created using stitched human videos, based on expert and deaf user feedback.

6 Significance of the Project with Respect to Research Contribution and Societal Benefits

This project marks a significant advancement in computer science by integrating three powerful fields: NLP, NMT, and computer vision. It is particularly impactful for low-resource languages like Sinhala, where existing tools and datasets are limited.

From a technical point of view, the project introduces innovative NLP techniques for both text and audio pre-processing. This includes developing robust tokenization and linguistic analysis methods tailored to Sinhala's complex morphology and grammatical structure. For audio inputs, speech-to-text models are employed to convert spoken Sinhala into written form, enabling seamless processing by downstream translation systems.

Building upon these foundations, the project utilizes NMT to convert processed Sinhala text, whether transcribed from speech or input directly, into accurate sign language gloss sequences. This approach allows the system to learn from the data and adapt to new sentence structures and expressions, surpassing the limitations of traditional rule-based translation systems. The application of deep learning in this context demonstrates its effectiveness in handling nuanced real-world language translation challenges.

From a societal perspective, this work has the potential to significantly improve accessibility for the deaf community in Sri Lanka. By providing an automated system that can translate Sinhala speech and text into structured sign language, it empowers people with deaf to participate more independently in education, public services, and daily communication.

Furthermore, the creation of a structured dataset and scalable AI-driven translation tools contribute to the preservation and standardization of SSL, supporting cultural and linguistic diversity while making the language more accessible for future generations.

In summary, this research advances key areas of computer science: NLP, speech processing, and machine learning, while delivering real-world social value through increased inclusion, accessibility, and equal opportunity for the deaf community.

7 Research Methodology

This research adopts the Design Science Research (DSR) methodology proposed by (Peffers et al. 2007), which provides a systematic approach for developing and evaluating innovative artifacts to solve real-world problems. The methodology follows a six-step process as illustrated in Figure 2, ensuring both rigour in the design of the research and relevance to practical needs.

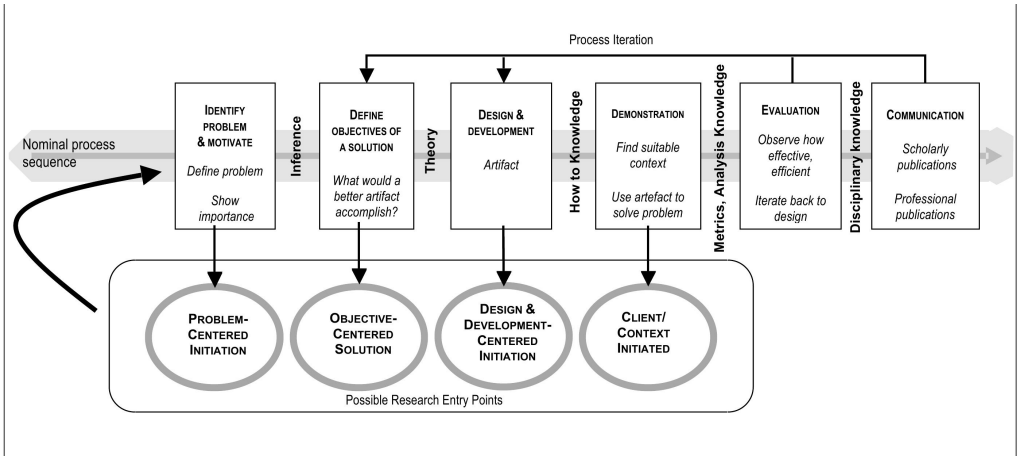

Figure 2: Design Science Approach

The DSR framework is particularly suitable for this study as it focuses on creating technological artifacts to address the practical problem of communication barriers faced by the Sri Lankan deaf community.

7.1 Step 1: Problem Identification and Motivation

As detailed in Section 1, this research addresses the communication barrier faced by Sri Lanka's deaf community due to the scarcity of SSL interpreters and limitations of existing rule-based translation systems. Problem identification was carried out through the following steps:

- Literature Review: Comprehensive analysis of global sign language translation systems (documented in section 3)
- Community Engagement: Collaboration with Moratuwa Deaf School to validate problem scope
- Gap Analysis: Identification of technical limitations in existing SSL tools

7.2 Step 2: Define Objectives of Solution

The solution objectives and research questions have been established in Section 5. The DSR approach will address these through the creation of multiple artifacts that collectively solve the identified problem.

7.3 Step 3: Design and Development

This step constitutes the core methodology for artifact creation and follows an iterative development approach.

7.3.1 Development Phase 1: Data Foundation

Artifact 1: SSL Video Dataset Creation

- Data Collection Protocol:
  - Collaborate with Dr. Rejinth Deaf School, Moratuwa, to record 1000+ SSL video clips.
  - Use controlled environment (consistent lighting, background, resolution)
  - Record common words, phrases, fingerspelling letters, and numbers.

7.3.2 Development Phase 2: Multimodal Translation

Artifact 2: Multimodal Translation Pipeline

- **Text/Audio Preprocessing**:

  The audio inputs will first be converted into text using Sinhala Automatic Speech Recognition (ASR) systems with integrated background noise reduction, thus improving the transcription accuracy.

- **Neural Machine Translation** : We propose a transformer-based neural translation model tailored for Sinhala grammatical structures, including tenses, cases, and inflections. The model will employ attention mechanisms to effectively capture contextual relationships and will be trained on a curated, well-documented dataset. To ensure real-time usability, we aim to design a low-weight model suitable for on-device (mobile) inference through optimization techniques such as model pruning and quantization.

7.3.3 Development Phase 2: Sign Synthesis Systems

Artifact 2:Human Video Synthesis

Previous Sri Lankan research on SSL used avatar- or skeleton-based representations, which are limited in conveying natural facial expressions and lip movements (Punchimudiyanse et al., 2015). To overcome this, our system uses pre-recorded human interpreter video clips. By stitching these clips with dynamic time warping, we generate fluent, naturalistic sentences that capture both

hand gestures and facial expressions, ensuring better clarity, expressiveness, and comprehensibility compared to avatars or skeleton animations.

7.4 Step 4: Demonstration

Proof-of-Concept Development: This stage involves building and showing a working version of my system that translates Sinhala speech or text into SSL videos, using real examples to demonstrate how it helps hearing people communicate with deaf individuals.

7.5 Step 5: Evaluation

The evaluation framework follows the metrics and user study design outlined below.

- **Technical Evaluation**: For this system, we will use BLEU-4 and Word Error Rate (WER) as technical metrics because they are standard and reliable for evaluating translation quality in neural machine translation, as supported by (Stoll et al. 2018) and others. This Bilingual Evaluation Understudy (BLEU) is especially used to evaluate automatic machine translation mechanisms (Chaudhary et al. 2022).BLEU-4 evaluates the performance of 4-gram words, i.e., four consecutive words, while BLEU-1 evaluates the individual word-based performance i.e. 1-gram. WER captures word order and substitution errors. These metrics are widely used in NMT research due to their objectivity and comparability across studies (Kahlon & Singh 2023). In addition, we will monitor the response time and data consumption to ensure that the system is efficient and practical for real-world deployment, as highlighted in (Dhanjal & Singh 2020).

- **User Evaluation**: We will conduct a comprehensive user evaluation involving two distinct participant groups: (i) Sign Language experts and (ii) Deaf students from the Moratuwa Deaf School. The evaluation aims to assess the linguistic quality, naturalness, and understandability of the SSL output generated by our system.

  - **Evaluation by Sign Language Experts:** Sign experts who are fluent in both Sinhala and SSL will evaluate the generated sign videos using a structured 5-point Likert-scale questionnaire. The evaluation will focus on linguistic and technical aspects of the translation. Each sentence in the test set will be assessed based on the following criteria:

    * Translation Correctness (Accuracy): Whether the generated SSL output accurately conveys the meaning of the original Sinhala sentence.
    * Grammar and Syntax: Whether SSL-specific grammatical structures (word order, topic markers, and sentence type) are correctly represented.
    * Expression and Facial Grammar: Appropriateness and accuracy of facial expressions and non-manual markers for conveying tone, emotion, and sentence type.
    * Flow and Transitions: Smoothness and naturalness of sign transitions, identifying any robotic or abrupt movements.

* Fingerspelling of Unknown Words: Clarity, timing, and recognizability of fingerspelling for out-of-vocabulary words such as names or numbers.
* Overall Naturalness: Overall impression of fluency and natural communication in SSL.

Each aspect will be rated on a 5-point scale (1–Very Poor to 5–Excellent). The mean and standard deviation of ratings will be computed, and qualitative comments from experts will be collected for iterative system refinement.

– **Evaluation by Deaf Students:** Deaf students from the Moratuwa Deaf School will participate in a comprehension-based evaluation facilitated by a sign teacher. The evaluation will focus on the understandability and usability of the generated SSL videos.
  * Students will watch the system-generated sign videos without subtitles and will be asked to explain what they understood.
  * The sign teacher will record their interpretations, which will then be compared with the original Sinhala sentences to measure comprehension accuracy.
  * Comprehension accuracy will be scored as follows: Correct (1 point), Partially Correct (0.5 point), Incorrect (0 points).
  * In addition, students will rate the video outputs on a simplified 5-point scale regarding the clarity of signs, naturalness of movement, facial expression effectiveness, and overall understandability.

  The comprehension accuracy percentage and mean user ratings will be calculated to assess the system's real-world performance and accessibility for the Deaf community.

– **Analysis and Feedback:** Quantitative results from both expert and Deaf user evaluations will be aggregated and analyzed. Qualitative feedback will be reviewed to guide future enhancements, particularly in improving gesture transitions, timing, and grammatical alignment.

- **System Usability Scale (SUS)**: To evaluate the overall usability and user experience of the system from the perspective of Deaf and hearing users. (Figure:3).

7.6 Step 6: Communication

The research results will be communicated through academic publications, open source releases, and community engagement.

8 Preliminary results and discussion

8.1 Data Gathering

The initial phase of this study focused on creating a foundational dataset essential for developing a Sinhala-to-Sinhala Sign Language (SSL) translation system.

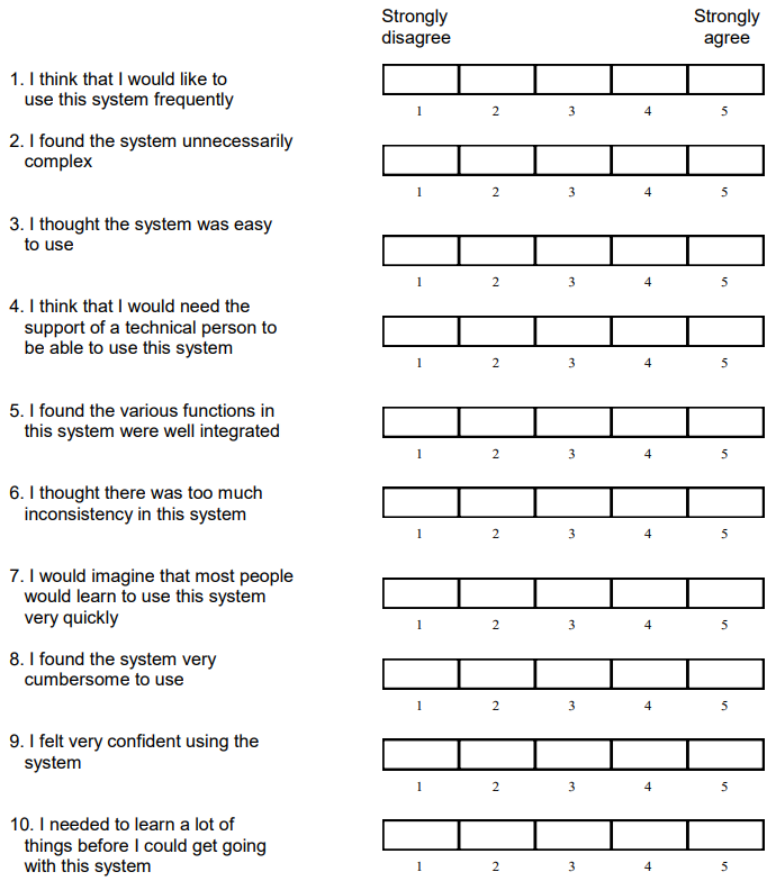

Figure 3: System Usability Scale (Brooke 1995)

8.1.1 Ethical Clearance and Permissions

Prior to data collection, the following approvals were obtained:

1. Permission from Dr. Reijinth, School of the Deaf, Moratuwa, to record sign language videos of students.
2. Ethical clearance from the University of Colombo School of Computing (UCSC) to ensure compliance with human-subject research guidelines, including informed consent from participants.

These steps were critical to ensure that the data collection adhered to ethical standards and respected participant privacy.

8.1.2 Video Data Collection

The dataset was collected from 1,208 individual videos, covering 1,208 unique words across multiple sign language categories, including nouns, verbs, adjectives, numbers, letters, and case-marked forms. The primary reference for this dataset was the Sri Lankan Sign Language (SSL) Dictionary (Figure 4, which consists of two volumes published by the Sri Lankan Central Federation of the

Figure 4: Sri Lankan Sign Language (SSL) Dictionary

Deaf. Together, these books contain around 4,000 signs covering commonly used words and are organized into 83 categories, such as Education, Health, Emotions, Food, and Nature. Each category provides distinct SSL signs with clear visual references, serving as a comprehensive guide for creating accurate sign representations. For the purposes of this project, 47 of the most commonly used categories were selected as an initial step, based on consultations with supervisors and experienced sign language teachers. During dataset creation, sign recordings were performed with the support of Deaf community interpreters, adhering closely to the dictionary's standards. In cases where the dictionary contained uncommon or rarely used signs, the teachers provided corrections and recommended the signs most commonly used in everyday communication. The videos are divided among each category, as visualized in Figure 5 and 6 below.

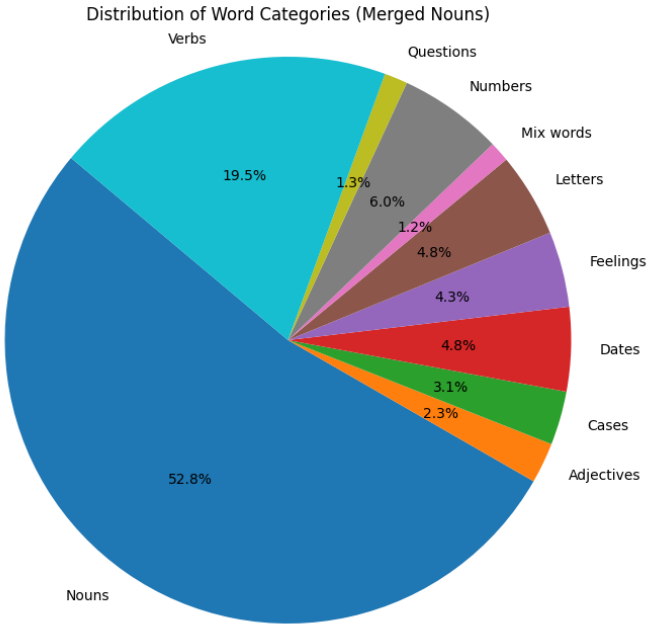

Figure 5: Distribution of word categories

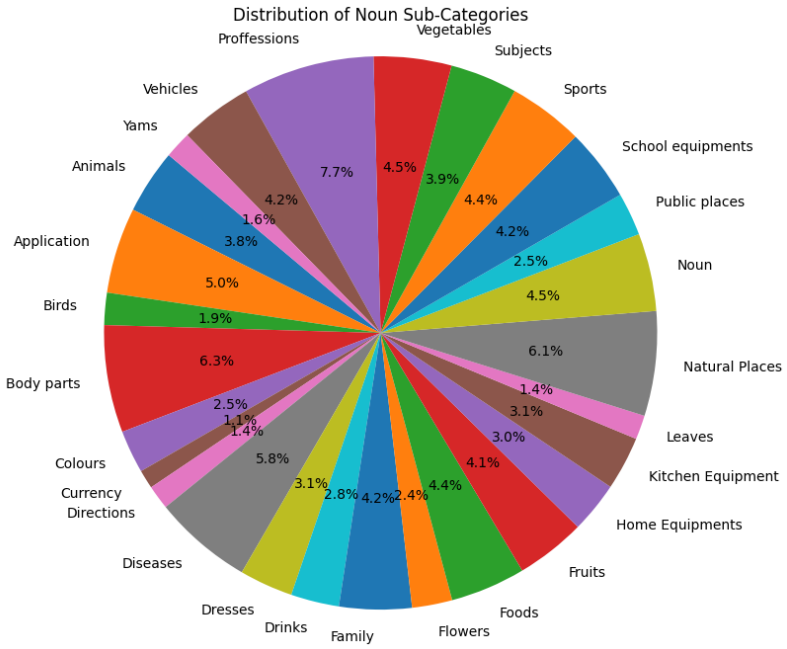

Figure 6: Distribution of noun subcategories.

Procedure:

The video dataset was recorded using an iPhone 11 at 4K resolution (30 fps) to ensure high visual quality. All sessions were conducted indoors under uniform lighting conditions. A plain white background and a blue outfit were used to provide a clear contrast with the signer's hand movements and facial expressions. Each word or phrase was recorded as an individual video clip, with multiple takes captured to select the most accurate representation. Camera framing was standardized to capture the upper body, ensuring visibility of both manual and non-manual features (Figure 8 and 7). The recorded clips were organized systematically by category, labeled with gloss and English/Sinhala meaning, and reviewed by sign language experts for accuracy and consistency.

8.2 Data Preparation and Gloss Generation

After collecting and categorizing all sign language video samples, the next step involved preparing the corresponding textual dataset required for translation model development. This phase focused on extracting Sinhala word labels, cleaning the dataset, and generating accurate English glosses to create a standardized bilingual mapping between Sinhala words and their sign language gloss representations.

8.2.1 Data Cleaning and Label Verification

All collected videos were first reviewed with the help of trained sign language teachers, and referring to the sign language dictionary ensured that each video represented the intended word correctly. The following steps were carried out:

- Removal of Incorrect or Partially Recorded Videos: Videos with incomplete signing, poor

Figure 7: On-site recording setup with lighting arrangement and guidance from teachers assisting the student performer.

lighting, or incorrect gestures were discarded under teacher supervision. A total of 312 videos were removed during this validation stage.

- Label Correction and Annotation: Each remaining video was assigned its corresponding Sinhala word label. Teachers verified the accuracy of these labels using standard sign language dictionaries and classroom teaching material.
- Data Cleaning : The dataset Excel file was filtered to remove,
  - Empty rows or cells without labels
  - Records referring to deleted or invalid videos
  - Duplicated Sinhala word entries

The resulting cleaned dataset contained 1208 unique Sinhala words across 36 categories.

8.2.2 Automated Gloss Generation

Once the cleaned Sinhala word list was finalized, a Python-based OpenAI gloss generation pipeline was developed to produce English gloss equivalents for each word.

- Load Cleaned Dataset

Figure 8: Close-up view of the video recording process during dataset preparation

- Context-Aware Gloss Generation: The model was prompted to produce ONE short English word or phrase (ALL UPPERCASE) reflecting the sign language meaning.
- Batch Processing and Saving: Dataset processed in batches of 10 rows. Generated glosses were saved to Excel after each batch for safety.
- Manual Verification: After automated generation, each English gloss was manually reviewed to ensure semantic accuracy and proper sign meaning alignment. Duplicate glosses (e.g., different Sinhala words mapping to the same English gloss) were identified and corrected. Literal translations that did not match the signing convention meanings were manually edited.

8.3 Parallel Corpus Generation

8.3.1 Experiment 01

To generate a high-quality Sinhala–Sign Gloss parallel dataset, it was essential to evaluate how effectively existing multilingual models capture semantic relationships in Sinhala. Since no gold-standard Sinhala–Gloss corpus currently exists, the aim of this phase was to identify the most semantically robust embedding method and tokenizer configuration suitable for dataset construction.

1. **Objective:** This experiment was designed to analyze how different multilingual models tokenize Sinhala text and to measure the semantic closeness between Sinhala words and potential

gloss representations. The results guided the selection of the most appropriate model for automatic gloss generation and alignment.

2. **Methodology:** Four embedding-based strategies were experimented with to analyze Sinhala tokenization and semantic similarity: **mBERT, SentenceTransformer (MiniLM), FastText, and mBART**. Each model was used to extract vector embeddings for a selected list of Sinhala words representing common nouns and verbs (e.g., "අම්මා" (mother),"ගුරු" (teacher),"පොත" (book), etc.). These embeddings were then compared based on cosine similarity and visualized using t-distributed Stochastic Neighbor Embedding (t-SNE).

   The experiment was conducted using Python and several open-source NLP libraries.

   - Transformers: Used for mBERT and mBART embeddings.
   - SentenceTransformers: Used for multilingual MiniLM embeddings.
   - Gensim: Used to load pre-trained FastText Sinhala embeddings (cc.si.300.vec).
   - Scikit-learn Seaborn: Used for similarity analysis and visualization.

   Each model generated vector representations for the same set of words. The cosine similarity, Euclidean similarity, and a hybrid similarity metric (average of cosine and Euclidean) were used to quantify semantic closeness.

3. **Results and Visualizations**

   Figures 9, 11, and 11 show the cosine similarity heatmaps for Sinhala words using four embedding models. The heatmaps illustrate pairwise semantic relationships between words, where darker shades represent higher similarity values.

   Among the evaluated models, SentenceTransformer (MiniLM) and mBERT demonstrated superior performance in capturing semantic relationships between Sinhala words. The FastText model, despite being based on subword information, exhibited slightly lower similarity consistency, possibly due to limited Sinhala corpus coverage during its training.

   The mBART model also performed reasonably well, reflecting its ability to encode multilingual sentence-level semantics. However, its performance was less stable for individual word-level representations since it was trained primarily for sequence-to-sequence tasks.

   These results highlight the importance of choosing context-aware models like mBERT or MiniLM for semantic-level tasks in Sinhala, whereas FastText remains effective for general lexical similarity computations.

   After generating gloss mappings using these models, several instances of inaccurate or irrelevant pairings were observed. Therefore, a manual validation process was introduced to refine and verify the dataset, ensuring higher semantic accuracy for downstream gloss-pair generation.

As a future direction, the approach will be extended from individual word-level embeddings to **sentence-based gloss evaluation**. This will allow a more context-aware assessment of meaning preservation between Sinhala sentences and their corresponding Sign Language gloss representations, providing deeper insight into semantic alignment for sign translation.

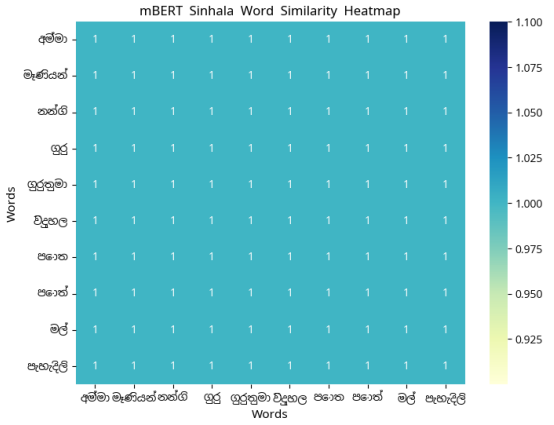

Figure 9: Cosine similarity heatmap for Sinhala words using BERT embeddings

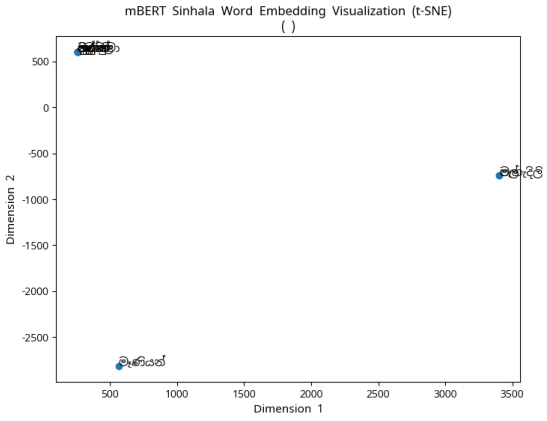

Figure 10: Spatial distribution of Sinhala word embeddings visualized using t-SNE in BERT

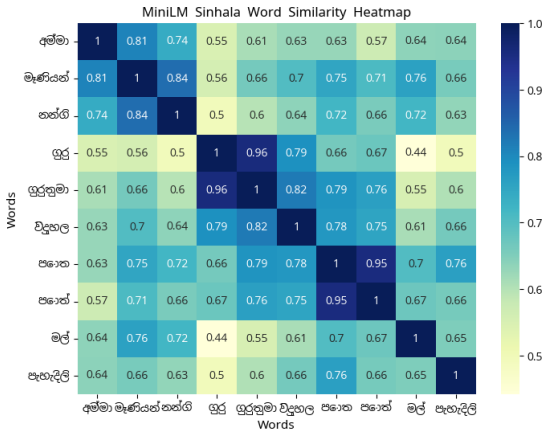

Figure 11: Cosine similarity heatmap for Sinhala words using multilingual MiniLM embeddings

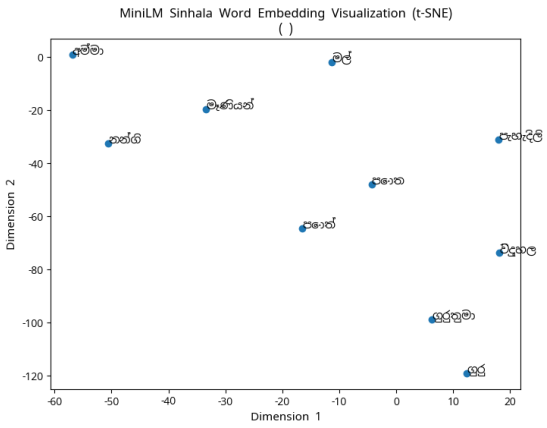

Figure 12: Spatial distribution of Sinhala word embeddings visualized using t-SNE in multilingual MiniLM

8.4 Baseline Model Development

8.4.1 Rule-based baseline

As an initial step toward automatic translation, we developed a rule-based baseline system inspired by the approach described in (Punchimudiyanse & Meegama 2017*a*) This rule-based method applies syntactic and morphological transformation rules to reorder Sinhala word sequences into gloss format. It serves as a baseline for comparing later neural approaches. The rules include:

- Removal of grammatical particles and auxiliary verbs,

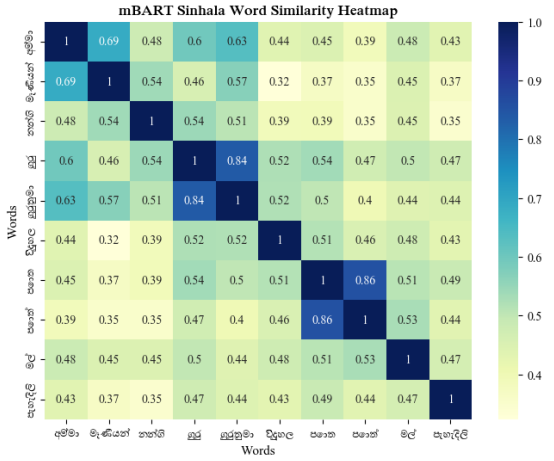

Figure 13: Cosine similarity heatmap for Sinhala words using multilingual mBART

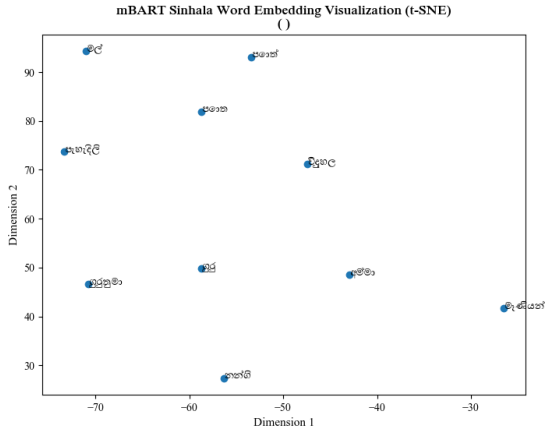

Figure 14: Spatial distribution of Sinhala word embeddings visualized using t-SNE in mBART

- Conversion of case markers into positional order, and
- Simplified word-order transformations to match sign syntax

8.4.2 Evaluation of Rule-Based Fingerspelling and Number Handling

This experiment focused on testing and validating the rule-based sub-module responsible for handling unknown words (fingerspelling) and numerical expressions, which was initially inspired by the Sinhala grammar-based transformation approach of (Punchimudiyanse & Meegama 2017*a*). The primary objective was to ensure that the module could accurately identify and process out-of-vocabulary (OOV) words and numbers during gloss generation and video synthesis.

Objective

To evaluate whether the rule-based method correctly identifies and applies fingerspelling for Sinhala words not found in the gloss dictionary, and whether number handling is accurate when Sinhala numerals or numeric text appear in the input sentence.

Methodology

A Python-based prototype was developed to automatically detect and handle unknown words and numbers in accordance with the proposed algorithm(rule-based approach). The number representation of this paper involves mapping input numbers (digits) to signal sounds (e.g., 234 = desiya thuna hathaxra). Each number corresponds to a presentation in the algorithm, which uses a recursive method.

Evaluation

There are 100 OOV words and 50 numbers used as a testing dataset. Two main metrics were used to evaluate the performance of the rule-based handling of fingerspelling and number representation:

- **Fingerspelling Handling Accuracy (FHA):**

$$\text{FHA} = \frac{\text{Correctly fingerspelled words}}{\text{Total OOV words}} \times 100 \tag{1}$$

- **Number Handling Accuracy (NHA):**

$$\text{NHA} = \frac{\text{Correctly represented numbers}}{\text{Total numeric tokens}} \times 100 \tag{2}$$

Results

Fingerspelling handling

The fingerspelling handling component achieved a 93% accuracy, successfully spelling out unknown Sinhala words by decomposing them into Unicode characters and retrieving corresponding SSL alphabet clips. The model handled proper nouns, borrowed words, and place names effectively using this method. However, that algorithm could not handle all the word creation using fingerspelling, because Sinhala has more modifiers that cannot be visualized using only limited rules. Nevertheless, this rule-based method is effective in handling fingerspelling according to the technical metrics of tokenizing and mapping, but still needs to be evaluated with final video outputs from sign experts, and we want to find an approach to connect this with our model creation to use a fallback method for handling out-of-vocabulary (OOV) words.Table 1 shows example words and their phonetic representations among the test dataset. From this, we can identify that some modifiers could not be identified using this approach, even though we annotate those letters (e.g., ඳ).

Numbering Representation

The module that was suggested by (Punchimudiyanse & Meegama 2017*a*) achieved 52% accuracy. The main limitation was identified during speech-to-text (STT) input, where numbers spoken as Sinhala text most of the time(e.g., "විස්ස" for 20) were not recognized as digits, resulting in incorrect mapping to number videos. This highlights the need for an improved number detection method — possibly using a Sinhala numeric lexicon (e.g., එක, දෙක, තුන, ...) to identify and map numbers expressed as text.

Additionally, preliminary feedback from Deaf community consultations (Moratuwa Deaf School) revealed that certain SSL number representations differ contextually. A clarification and formalization step will be conducted with sign language experts to finalize these numeric conventions. Also, the current method does not match our dataset and approach. So, we are planning to implement a method to identify the textual representation of numbers, convert them into digits, and subsequently find the correct sign video for each number. In the updated version, we implemented a method to handle numbers in both digit and textual forms, ensuring accurate mapping to our dataset of Sinhala Sign Language (SSL) number videos. The process involves three main steps:

1. **Textual-to-Numeric Conversion:** Numbers written in Sinhala words (e.g., හයසිය හතලිහ

Table 1: Fingerspelling Evaluation Results

Word	**Phonetic Representation**	**Incorrect Letters**	**Correctly Identified Count**
දිනූෂා	dh i n uxu zsh axa	-	6
තීරුව	th ixi r u v a	-	6
විද්‍යාලය	v i dh wqx y axa l a y a	ද්‍යා	7
ගොඩනැගිල්ල	g o d a n ae g i l w la	-	10
වෙළඳසැල	v e L a ඳ s ae l a	ඳ	8
රවීන්ද්‍ර	r a v ixi n w dh w qx r a	ද්‍ර	4
නිරෝෂා	n i r oxo zsh axa	-	6
කවිඳු	k a v i ඳ u	ඳු	5
අපේක්ෂා	a p exe k w zsh axa	-	6
ගම්පහ	g a mw p a h a	-	7
කුරුණෑගල	k u r u N ae xa g a l a	-	11
අනුරාධපුර	a n u r axa zdh a p u r a	-	11
මාතර	m axa th a r a	-	6
මාවනැල්ල	m axa v a n ae l wla	-	9

හත for 647) are first converted into their corresponding numeric values using a dictionary-based mapping. This ensures that the textual representation is correctly interpreted as a numerical value.

2. **Dataset-Aware Decomposition:** The numeric value is then decomposed into subcomponents that match the available sign videos in our dataset. For example, 647 is decomposed into 600, 40, and 7, corresponding to the signs හයසිය, හතලිහ, හත.

3. **Sign Sequence Generation:** Finally, each decomposed numeric component is mapped to its respective sign video. This ensures that the avatar plays the correct sequence of gestures corresponding to the input number, maintaining the natural and dataset-consistent animation (Table **??**).

This approach allows flexible handling of input numbers in various forms and guarantees that every number can be accurately converted to a sign sequence within the constraints of our SSL number video dataset.

Table 2: Examples of Updated Number-to-Sign Conversion Results

Input	Numeric Value	Sign Sequence	Correctness
0	0	0	Correct
7	7	7	Correct
15	15	15	Correct
29	29	29	Correct
42	42	40, 2	Correct
58	58	50, 8	Correct
73	73	70, 3	Correct
86	86	80, 6	Correct
99	99	90, 9	Correct
100	100	100	Correct
256	256	200, 50, 6	Correct
512	512	500, 12	Correct
999	999	900, 90, 9	Correct
1000	1000	1000	Correct
2345	2345	2000, 300, 40, 5	Correct
6789	6789	6000, 700, 80, 9	Correct
දෙදහස පහ	2005	2000, 5	Correct
හතරදහස නවය	4009	4000, 9	Correct
තුන්දහස හතර පහ	3009	3000, 9	Incorrect
එක්දහස එකසිය පහ	1105	1000, 100, 5	Correct
නවදහස නවසිය නවය	9909	9000, 900, 9	Correct

8.5 Experiment 2: Video Synthesis for Sinhala Fingerspelling

8.5.1 Background:

To validate the feasibility of generating Sinhala sign language videos, we conducted an experiment focused on fingerspelling and word-level video synthesis. The goal was to test whether individual video clips for letters, vowels, modifiers, and numbers could be concatenated to form accurate, continuous sign language representations of Sinhala words. This experiment aimed to verify the correctness of phonetic-to-video mapping and the visual intelligibility of synthesized sequences.

8.5.2 Methodology:

- **Dataset Preparation:** A flat video repository (videos/) was created containing clips for single letters, vowels, numbers, and common syllables. File names were mapped to a phonetic tag representation derived from the Sinhala text.
- **Text-to-Video Pipeline:** Input Sinhala words were first converted into phonetic sequences using a custom Unicode-to-phonetic converter.Phonetic tags were then matched against the video dataset.Corresponding clips were concatenated in order using **MoviePy** to produce a final video representing the full word.
- **Testing Procedure**: Words of varying lengths (5–15 characters) were tested, including proper nouns and daily vocabulary.

- **Results and Observations:**
  - For shorter words (<5 clips), the generated video sequences correctly represented the intended Sinhala fingerspelling. Example: "දිනූෂා" produced the expected sequence of hand gestures corresponding to each letter.
  - Issues:
    - Memory Limitations: Longer words caused memory overflow errors during concatenation, even after reducing video resolution (640px width) and compression (CRF = 30).
    - Processing Time: Concatenation of multiple clips (especially >20) took significant time due to in-memory loading of all clips.
    - Dataset Coverage: Some phonetic tags were missing corresponding videos, resulting in incomplete sequences.
  - Reasons for the Issues: MoviePy loads all video clips into RAM before concatenation, making long sequences memory-intensive. High-resolution clips increase RAM usage, which exacerbates memory overflow. Lack of optimized streaming or pre-processing increases both memory and processing load.
  - The experiment confirmed that video-based sign synthesis is feasible for Sinhala, but practical limitations in memory and processing require further optimization.

9 Remaining work

The next phase of this research will focus on improving the translation accuracy, optimizing model performance, and implementing an end-to-end multimodal Sinhala-to-Sri Lankan Sign Language translation pipeline. The remaining tasks are outlined as follows:

9.1 Dataset Enhancement and Gloss Mapping

Although a semi-automated gloss generation pipeline has been developed using transformer-based multilingual embedding models, it was observed that the semantic accuracy of some generated gloss pairs is limited due to the linguistic complexity of Sinhala. To address this, the next step will focus on refining the dataset through a semi-automated and manual hybrid process:

- Conduct manual validation and correction of Sinhala–Gloss pairs with the guidance of sign language teachers from the Moratuwa Deaf School.
- Generate a high-quality Sinhala-to-Gloss parallel dataset by combining automated outputs with expert-reviewed samples.
- If the semi-automated method continues to yield low alignment accuracy, a fully manual dataset creation process will be initiated to ensure linguistic and semantic reliability.

- Expand the dataset to include complex sentences, not just isolated words, ensuring proper syntactic structures aligned with SSL grammar.

9.2 Model Development and Optimization

The next phase will involve the design and implementation of a transformer-based translation model to convert Sinhala text into sign gloss sequences:

The curated parallel Sinhala–Sign Language dataset will be used to train a transformer-based Neural Machine Translation (NMT) model using PyTorch, focusing on a lightweight architecture that balances translation quality with computational efficiency to support mobile deployment. Alternative transformer models, such as DistilBERT and MarianMT, will be evaluated to select the one offering optimal performance under limited resource conditions. The model will be fine-tuned using SSL-specific linguistic constraints, and performance will be assessed using standard metrics like BLEU-4 and Word Error Rate (WER). However, as the model alone cannot fully handle out-of-vocabulary words, proper names, and numeric expressions, an additional module will be integrated to address these cases. Specifically, rule-based fingerspelling and number representation methods—developed and experimentally validated in prior work—will be incorporated to complement the NMT outputs, ensuring that unknown words and numbers are accurately conveyed in SSL videos. This hybrid approach allows the system to maintain both linguistic correctness and naturalistic video representation for real-world communication.

9.3 Sign Gesture Generation (Sign Synthesis Phase)

Once the model outputs the translated gloss sequence, the next step is to generate the corresponding SSL video output:

- Retrieve pre-recorded SSL video clips corresponding to each gloss token.
- Optimize video storage and retrieval mechanisms, as the current 4K dataset consumes substantial memory and slows concatenation.
- Explore compressed video formats and cloud-based storage (e.g., Firebase Storage, AWS S3) with on-demand loading to reduce device resource consumption.
- Implement an efficient video stitching pipeline using dynamic time warping (DTW) to ensure smooth transitions and natural movement.
- Conduct experiments to reduce frame redundancy while maintaining sign clarity.

9.4 Evaluation and Validation

Comprehensive evaluation will be conducted to assess the system's performance and usability. Ongoing experimental evaluations of model configurations will continue to refine accuracy and fluency. Additionally, user evaluations involving sign language experts and deaf students will measure

key qualitative factors such as understandability, naturalness, and expressiveness. Quantitative metrics such as BLEU-4 and Word Error Rate (WER) will be used to evaluate translation performance, while Likert-scale ratings will capture human judgments. To assess practical usability, System Usability Scale (SUS) testing will be carried out with both deaf and hearing participants, ensuring that the system meets real-world communication needs.

9.5 Documentation and Dissemination

All findings, evaluations, and system outcomes will be documented for academic publication in venues specializing in accessible technologies and sign language processing. To encourage future research, cleaned datasets, baseline models, and evaluation scripts will be released as open-source resources.

10 Timeline

This study is expected to last 16 months, and Table 3 shows the planned timeline and activities for the study.

Table 3: Expected Project Timeline

Activity	**Jan - May 2025**	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr
Problem Identification and Literature Review												
Formulation of Research Objectives and Questions												
Video Dataset Creation												
Development of Neural Machine Translation Model												
Sign Synthesis via Human Video Stitching												
Evaluating Results												
Iterative Refinement												
Dissemination and Documentation												

References

Ahinsa, P., Thrimahavithana, S. & Karunanayaka, K. (2025), 'Bridging communication gaps: Advancements, challenges, and future directions in text-to-sign language translation', *Journal of Future Artificial Intelligence and Technologies* **2**(1), 110–134.
URL: *https://faith.futuretechsci.org/index.php/FAITH/article/view/91*

Alaghband, M., Maghroor, H. R. & Garibay, I. (2023), 'A survey on sign language literature', *Machine Learning with Applications* **14**, 100504.

Bertin-Lemée, E., Braffort, A., Challant, C., Danet, C. & Filhol, M. (2023), Example-based machine translation from textto a hierarchical representation of sign language, *in* M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. A. Vidal, N. Aranberri, M. Nunziatini, C. P. Escartín, M. Forcada, M. Popovic, C. Scarton & H. Moniz, eds, 'Proceedings of the 24th Annual Conference of the European Association for Machine Translation', European Association for Machine Translation, Tampere, Finland, pp. 21–30.
URL: *https://aclanthology.org/2023.eamt-1.3/*

Brooke, J. (1995), 'Sus: A quick and dirty usability scale', *Usability Eval. Ind.* **189**.

Brour, M. & Benabbou, A. (2019), Atlaslang mts 1: Arabic text language into arabic sign language machine translation system, *in* 'Procedia Computer Science', Vol. 148, Elsevier B.V., pp. 236–245.

Chaudhary, L., Ananthanarayana, T., Hoq, E., Nwogu, I. & Member, S. (2022), 'Signnet ii: A transformer-based two-way sign language translation model'.
URL: *https://www.ieee.org/publications/rights/index.html*

Daily Mirror (2018), 'Sri lanka terribly short of sign language interpreters'. Accessed: 2025-06-07.
URL: *https://www.dailymirror.lk/News-Features/Sri-Lanka-Terribly-short-of-sign-language-interpreters/131-147833*

Delorme, M., Filhol, M. & Braffort, A. (2009), Animation generation process for sign language synthesis, *in* 'Proceedings of the 2nd International Conferences on Advances in Computer-Human Interactions, ACHI 2009', pp. 386–390.

Dhanjal, A. S. & Singh, W. (2020), 'An automatic conversion of punjabi text to indian sign language', *EAI Endorsed Transactions on Scalable Information Systems* **7**, 1–10.

Fernando, P. & Wimalaratne, P. (2016), 'Sign language translation approach to sinhalese language', *GSTF Journal on Computing (JoC)* **5**. Udugama LSK, Nethsinghe R, Southcott J, Kularathna S, Dhanapala TDTL, Alwis KAC. Sign language usage of deaf or hard of hearing Sri Lankans. J Deaf Stud Deaf Educ. 2024 Mar 17;29(2):187-198. doi: 10.1093/deafed/enad055. PMID: 38073324.

Ghosh, A. & Mamidi, R. (n.d.), 'English to indian sign language: Rule-based translation system along with multi-word expressions and synonym substitution'.

Groundviews (n.d.), 'Learning sri lankan sign language'. Accessed: 2025-01-21.
URL: *https://groundviews.org/2021/09/02/learning-sri-lankan-sign-language/*

Gupta, V., Sinha, S., Bhushan, P. & Shettigar, M. (n.d.), 'English text to indian sign language translator coen 296b-natural language processing'.

Kahlon, N. K. & Singh, W. (2023), 'Machine translation from text to sign language: a systematic review', *Universal Access in the Information Society* **22**, 1–35.

Martino, J. M. D., Silva, I. R., Marques, J. G. T., Martins, A. C., Poeta, E. T., Christinele, D. S. & Campo, J. P. A. F. (2023), 'Neural machine translation from text to sign language', *Universal Access in the Information Society* .

Morrissey, S. & Way, A. (2009), 'An example-based approach to translating sign language', *Morrissey, Sara and Way, Andy (2005) An example-based approach to translating sign language. In: Second Workshop on Example-based Machine Translation, 16 September 2005, Phuket, Thailand.* .

Munasinghe, N., Jayalal, S. & Wijayasiriwardhane, T. (2023), 'Interpretation of sri lankan sign language: A wearable sensor-based approach', *Proceedings - International Research Conference on Smart Computing and Systems Engineering, SCSE 2023* .

Patel, P. K. D., Vaghasiya, K. & Savaliya, R. (2025), 'Sign language translator with speech recognition integration: Bridging the communication gap', *International Journal of Scientific Research in Science, Engineering and Technology* **12**, 741–749.
URL: *https://www.ijsrset.com/index.php/home/article/view/IJSRSET25122201*

Peffers, K., Tuunanen, T., Rothenberger, M. & Chatterjee, S. (2007), 'A design science research methodology for information systems research', *Journal of Management Information Systems* **24**, 45–77.

Punchimudiyanse, M. & Meegama, R. G. N. (2015), 3d signing avatar for sinhala sign language, *in* '2015 IEEE 10th International Conference on Industrial and Information Systems, ICIIS 2015, Dec. 18-20,2015, Sri Lanka', IEEE, p. 551.

Punchimudiyanse, M. & Meegama, R. G. N. (2017*a*), 'Animation of fingerspelled words and number signs of the sinhala sign language', *ACM Transactions on Asian and Low-Resource Language Information Processing* **16**.

Punchimudiyanse, M. & Meegama, R. G. N. (2017*b*), 'Computer interpreter for translating written sinhala to sinhala sign language', *Journal* **12**, 70–90.

Rao, P. S., Rohit, T. V., Manasa, B. & Lingamaiah, P. V. (2023), 'Multiple languages to sign language using nltk', *International Journal of Scientific Research in Science and Technology* pp. 12–17.

Sewwantha, R. D. R. & Ginige, T. N. D. S. (2021), 'Mask region-based convolutional neural networks (r-cnn) for sinhala sign language to text conversion', pp. 207–219.

Stoll, S., Camgoz, N. C., Hadfield, S. & Bowden, R. (2018), Sign language production using neural machine translation and generative adversarial networks, *in* '29th British Machine Vision Conference (BMVC 2018) (Northumbria University, Newcastle Upon Tyne, UK, 03/09/2018–06/09/2018)', British Machine Vision Association.

Thrimahavithanaa, S., Yasodha, V., Kannangara, M. U., Welgama, V. & Weerasinghe, A. (2019), *19th International Conference on Advances in ICT for Emerging Regions (ICTer) - 2019 : conference proceedings : 03rd and 04th of September 2019, Vidya Jyothi Professor V K Samaranayake Auditorium, University of Colombo School of Computing, Colombo, Sri Lanka*, IEEE.

Tmar, Z., Othman, A. & Jemni, M. (2013), A rule-based approach for building an artificial english-asl corpus, *in* '2013 International Conference on Electrical Engineering and Software Applications, ICEESA 2013'.

Webd, U. & Grieve-Smith, A. (2001), 'Signsynth: A sign language synthesis application using web3d and perl'.