Tamil Lang TSP: Tamil Lang Transformer Neural Text to Sign Production

Thillai Sivakavi S
Department of Computing Technology, SRM Institute of Science and Technology Chengalpattu, India
thillaisivakavi75@gmail.com

Minu R I
Department of Computing Technology, SRM Institute of Science and Technology, Chengalpattu, India
minur@srmist.edu.in

Abstract: *Tamil lang Task-Specific Prompts (TSP) is an advanced machine translation system that seamlessly converts Tamil text into Tamil Sign Language. This innovative system integrates cutting-edge neural machine translation and motion graph technology to automatically generate sign language from the input text. The process involves a meticulous analysis of the morpho-syntactic structure of the Tamil sentence, followed by its conversion into American Sign Language (ASL) notation. This notation generates gloss, serving as a pivotal element for constructing a motion graph. The motion graph is then utilized to create pose sequences that align with the generated gloss. This pioneering approach represents the first complete pipeline for accurately translating Tamil language text into corresponding sign sequences. To evaluate its translation capabilities, this approach undergoes both quantitative and qualitative assessments using a custom-built dataset. Furthermore, its performance is compared with a German language translation system, providing valuable insights into its effectiveness.*

Keywords: *Sign language production, Roberta morpho syntactic analysis, sign language production network.*

Received July 4, 2024; accepted October 7, 2024
https://doi.org/10.34028/iajit/21/6/15

1. Introduction

Every person's life has been profoundly impacted by technological advances in various sectors. Neural machine translation is one such technical development that makes it easier for people to grasp a wide range of natural languages. Users can do neural translation from one natural language to another using platforms offered by Google and Microsoft [16]. Humans can now travel and read literature in various languages with ease because of these translation systems, and their main benefit is that they are readily available and affordable [7]. But these benefits aren't available to some deaf and dumb individuals. The primary cause of this is because their language is unique and unfamiliar to others. By 2050, the number of individuals with hearing loss is expected to rise to 700 million, or one in ten people, from the 5% of the global population currently projected to have hearing loss, according to the WHO [40]. To accommodate this community into the global community, it becomes unavoidable to express everything in their language as well. Translators are the obvious solutions to make the lives of this community easier. However, there are certain potential difficulties associated with developing such translators, as the language used by these people is sign language. Sign language utilizes the manual and non-manual components of a deaf person to express themselves. The manual components are primarily their hands and their shape, orientation, position, and movements, and under the non-manual component, their eye gaze, facial expressions, and postures are considered [9]. It is like any other language; sign language also has its own grammatical rules and linguistic structures. So, sign-to-text and text-to-sign translation is not a simple process of mapping a single word to every gesture [26]. As shown in Figure 1, the sentence "A frosty cold night before us" does not have a direct sign form, so this speech is converted to its corresponding text, and from there, an intermediate representation like Glove is used for mapping the actual sentence to its corresponding signs.

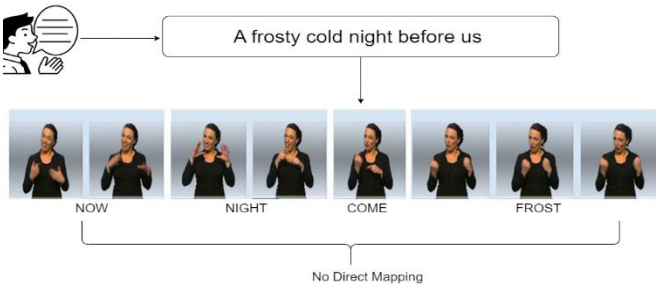


Figure 1. Translation of any language sentence to its corresponding sign language is not direct mapping and it always requires an intermediate form of representation for carrying out the mapping between the actual language and sign language.

Furthermore, sign language is not universal; it differs according to the language's syntax and vocabulary. Therefore, spoken language must be taken into account while designing machine translation systems. Because of this, it is also clear that an English native cannot comprehend Indian Sign Language (ISL) and vice versa. The talks are unique to the ISL because the focus of this effort is developing a translation system for the Tamil

language. These factors need the use of machine translation systems that can both translate spoken words into sign language and enable simple and effective communication between the deaf population and speakers of all other languages. This paper aims to investigate this. In this effort, we essentially take a look at a sentence written in Tamil and try to develop a strong system that can translate Tamil text into the appropriate sign.

1.1. Challenges

The majority of commercial applications center on Sign Language Recognition Systems (SLRs), which attempt to associate spoken language words with their appropriate signs. Recent research on sign language recognition in a variety of languages, including the ISL, is presented in the literature from [2, 3, 4, 11, 12, 32, 33, 41]. These indicate that there are misunderstandings about the deaf population in some capacity. People who are deaf or dumb are usually thought to be quite familiar with spoken language and do not need sign language translation. However, this could not be the case because writing and speaking a language vary. Another challenge is translating spoken language into sign language. This suggests that it is also necessary to construct sign languages for the deaf and dumb community to benefit from actual benefits enjoyed by another group, such as a SLR.

The process of producing a sign involves translating text to the appropriate sign language, and the avatar-based technique, as seen in [10], often supports this sign creation system. Previous avatar-based methods attempt to display the text's signals; nevertheless, the prepared sentences play a major role in this method's effectiveness [22]. A further popular method is translating spoken language into its associated sign gloss, which is subsequently mapped to the avatar as a parametric value that drives the avatar's movement. Nevertheless, as Figure 1 illustrates, the gloss representation in this method does not correspond to the total amount of words in the original language. Although these methods accomplish translation, machine translation may have issues with contextual translation.

1.2. Our Contributions

Getting the deaf and dumb people to manage everyday tasks independently is crucial to improving their quality of life and integrating them into society at large. Those who are deaf or dumb have challenges when it comes to things like watching television and understanding information posted on public notice boards. Their lives will thus improve if the same knowledge is also revealed in the form of signals. This can only be accomplished by converting text or speech to signs using a sophisticated sign-generating system. To advance the field of SLP and develop a first-of-a-kind SLP system for one of the ISLs, Tamil, a new approach is proposed in this work, harnessing methods from neural machine translation and neural networks. The proposed method inputs the Tamil language sentence in the form of text and outputs a sign language video. The transformer encoder and decoder network are used for the generation of gloss, which is then used to arrive at a pose sequence, which is eventually shown as video sequences.

Thus, the major contributions of this work are as follows:

- A transformer-based neural machine translation framework that converts the Tamil text to Tamil Sign language is built.
- The complete pipeline essential to building an SLP, from the process of capture to that of evaluation, is presented.
- To our knowledge, this is the first SLP system developed for the Tamil language deaf and dumb community.

The creation of this SLP will be very helpful in raising the technical proficiency of individuals with exceptional needs. The work presented here draws inspiration from [16], where we employed a transformer network as a neural machine translation system and found examples of its application in Tamil, in addition to demonstrating the usefulness of motion graphs for avatar motions. Comprehensive quantitative and qualitative assessments are conducted to appraise the suggested methodology.

The rest of this paper is organized as below:

Section 2 gives an overview of some of the advancements made in SLP and the usage of neural machine translation in SLP. Section 3 details the framework and the pipeline used for building the Tamil-language SLP system, followed by the experimentation and its evaluation. Section 5 discusses the limitations and concludes this work.

2. Background-Related Works

Sign Language Production (SLP) as mentioned earlier takes the other language text or speech as input and the output is sign. This process involves a translation based on neural networks and the output needs to be shown visually. Hence in this literature survey, we will look into some of the neural machine translation methods and the avatar technology to view the output in visual form. Additionally, we will understand the suitability of using the motion graphs for making the avatars show the signs.

2.1. Neural Machine Translation(NMT)

Neural machine translation systems that converted spoken language to sign started to emerge in the 1990s. Some significant works in several languages were created during this time. Translated from English to

another sign language [21], from Japanese-to-Japanese sign language [35], from English to American Sign Language (ASL) [36], and from English to ISL [23], ZARDOZ was translated. The text-to-sign translation has been carried out in recent years utilizing various machine-learning techniques. These methods can be broadly classified as follows:

1. Machine translation based on rules.
2. Segmented corpus-based translations, such as statistical and hybrid machine translation.
3. Machine translation using neural networks.

Given that neural machine translation is the foundation of our work, some of the research utilizing NMT is reviewed.

Neural machine translation uses the artificial neural network to make predictions regarding the next word in a sequence and also makes evaluations on the likelihood values of the words, and this is done by coding a sequence of words as a fixed-length vector [38]. The NMT methods for performing the translation from text to sign are still unexplored in various contexts, but a few significant works contribute to advancing this neural machine translation. One such work using the ASLG-PC dataset [43] was used to obtain the glosses for a given sentence and is presented in [7]. Though this approach performed the translation, there were a high number of tokenization errors because the glosses weren't annotated and the size of the vocabulary was small.

Arabic language to Arabic sign language translation is done in [7] where the problem solving is done using multiple steps in which the sentence encoding is done first with the guidance from the morphological characteristics of the words. The sign generated is shown as a 3D avatar which is guided by the SiGML coding. The target network is generated using the feed-forward neural network and this method generated a BLEU score of 0.79. Text2Sign [34] is another form of neural machine translation system in which the text is converted to sign with the assistance of the motion graphs without using the 3D avatar assisted by the Generative Adversarial Network. Recurrent Neural Network (RNN) and motion graphs together are used for the creation of pose sequences. Then these pose sequences are used for conditioning the GAN network that provides the output in the form of video sequence and this is first-of-a-kind work without using 3D avatar visualizations. This Text2Sign has shown significant results with varied datasets like RWTH-PHOENIX-Weather [27], SMILE [24], and Dynamic Visual Content from Broadcast Footage (Dynavis) [18]. The Text2Sign though robust suffers from low translation training and thus this approach couldn't compete with the avatar-based approaches.

A direct mapping between the sign language and poses is achieved in [13] without any intermediate form of glosses as done in Text2Sign. Here a progressive transformer architecture guides the transformer to learn the sentence as a sign pose sequence and thus the given input sequences are shown as signs in the skeleton pose. A similar approach of using the skeleton poses for the translation is proposed in [5].

The approaches mentioned use glosses or skeleton poses for performing the translation of the text to sign. Though these approaches perform translation, there is always a question of whether the skeleton-based poses are really useful for the deaf and dumb community. In this context, presenting the skeleton pose sequences in the form of videos or 3D avatars became a mandate, and a similar effort was made in [30] where it proved the importance of video-based visualization compared to that of the skeletal visualizations and it also enhanced for the photo-realistic visualization in [42].

These are some of the state-of-the-art NMT approaches and most of the work is based upon the PHOENIX-14 T dataset. From these methods, it is evident that sign synthesis using video or 3D avatars is essential for the generation of SLP system. So, in the next section let us review some of the approaches based on avatars.

2.2. Avatar Approaches for SLP

Avatars are essential in the SLP system as the text in natural language needs to be shown in sign language. For making the avatar understand the text format motion capture data and parameterized glosses are used. So as far as the SLP using avatars, several research projects have been carried out to make the avatars show better signs using parametrized glosses. State-of-the-art analysis shows that the avatar animations using parametrized glosses started in 2000. Visicast [39], Esign [31], Tessa [15], dicta-sign [44], and JASigning [8] are some of the SLP-similar projects that were able to produce signs when they are annotated using HamNoSys, DRS, HPSG and SigMLnotations. Though they served as a starting point for SLP they do not include the non-manual components and hence the popularity of these models among the deaf community is nil. Thus, some of the recent works incorporate the non-manual components as well for the avatar animations. The work presented in [14] generates rules for non-manual movements like the arm. The rules are supported by the geometric constructions using which the expressions of the Avatar are controlled. Since it involves the usage of production rules that help in matching the linguistic components to that of the expression, this guided avatar for producing the signs. Similarly, in [1] a specific system meant for the spine movements is proposed.

Motion capture data are then used for the movement of the avatars. This is presented in [17, 25] termed as Sign3DprojectbyMocapLab which moves the avatars using the motion capture data. Though the results based on the motion capture data are highly realistic it suffers

with the smaller phrases set and the uncanny valley also remains as a problem. Thus, to achieve a scalable avatar-based signing more realistic, the proposed framework using the transformer network is benefic.

3. Tamil Lang Transformer Text to Sign Translator Overview

This text-to-sign translation is expected to perform the process of translating the Tamil sentence into poses showing the signs of those sentences. To achieve this, we are using a morpho-syntactic analysis approach so that each word in the sentence is understood and the necessary signs can thus be achieved. So basically, in the morpho-syntactic analysis, we take each word of the sentence and construct a knowledge graph that helps us understand the context of the sentence. This step helps the avatar make sensible signs as the morpho-syntactic analysis results are mapped to the results of the avatar. Thus, the morpho-analyzed sentence is fed into the BERT transformer to carry out the process of encoding. This encoded vector then passes through a feed-forward neural network with backpropagation to achieve the neural machine translation. Then the knowledge-guided decoder helps the avatar make the movements. Let's discuss every part of the system in a detailed manner. Figure 2 presents the overall framework of the text-to-sign translation.

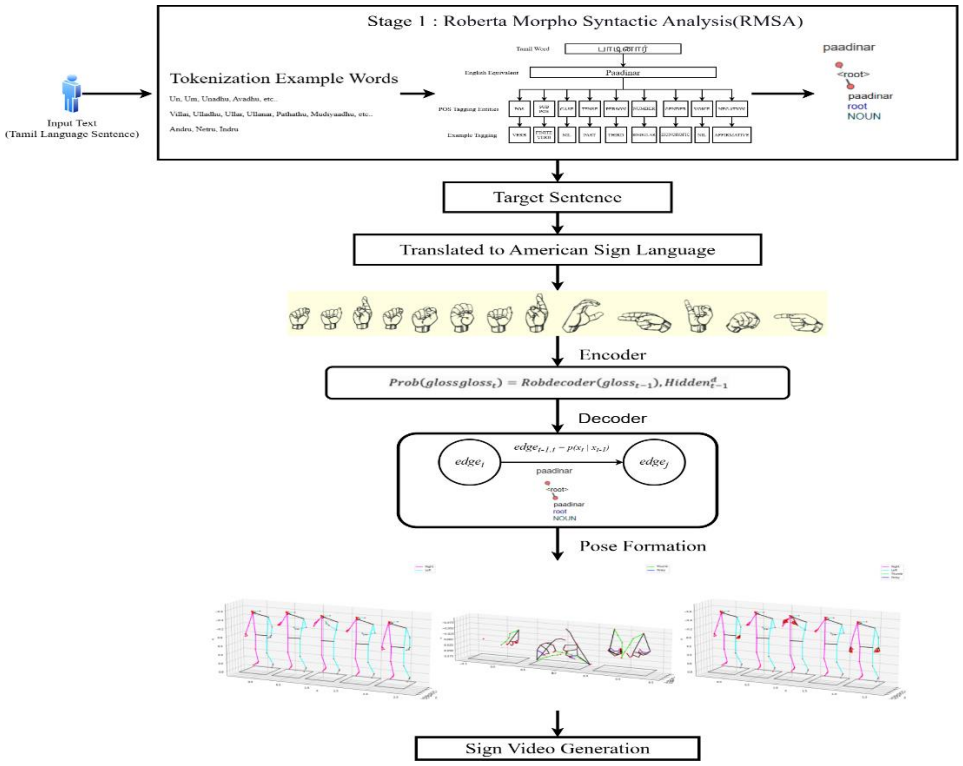


Figure 2. Tamil text to sign translation overall architecture.

3.1. Stage 1 Roberta Morpho-Syntactic Analysis (RMSA)

This part of the procedure is carried out because creating signals for words alone can assist in mitigating the issue of movement epenthesis by identifying the significance of each word in a phrase and its context. To address the syntactic analysis problem, a transformer network is employed to describe the problem as a sequence labeling problem.

3.1.1. Problem Formulation

The objective of this phase is to identify the most important words in the sentence. Thus, this problem can be modeled as a sequence labeling problem as the general definition of the sequence learning problem takes the below form.

Given a spoken Tamil language sentence S_n={w_1, w_2, w_3… … … .. w_n} as input, where w_1, w_2, $w3$… … … .. w_n represents the individual words and n denotes the length of the sentence. The output is also a sequence of words represented as O={O_1,O_2,O_3… … … .. O_n,} and the O is framed by labeling each of the words in S_n. The labeling here denotes the syntactic association to each word of the sentence. The syntactic association here denotes the POS tagging, noun chunking, and named entity recognition.

3.1.2. ROBERTa Model

For solving the problem formulated in section 3.1.1. transformer Roberta [19] is used for mapping the sequence of words to its corresponding labels by modeling it as a conditional probability $P\ S_n|O$. The input sentence in Tamil language is analyzed using the improved version of the tree bank parser [37]. The

Morpho-syntactic representation of each word in the sentence is obtained after it undergoes the process of tokenization, Part Of Speech (POS) mapping, systematic changes, and dependency relation mapping. Figure 3 shows the RMSA process.

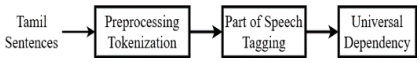


Figure 3. Roberta Morpho syntactic analysis annotation process.

- **Preprocessing**

We haven't taken into account any raw corpus because the goal of this method is to translate basic statements into their appropriate signs; thus, transliteration of the sentences wasn't necessary. Tokenization was performed as a first preprocessing step, using spaces as the primary means of word separation in the phrases. In addition, Tamil words frequently mix nouns with postpositions and auxiliary verbs with lexical verbs, in contrast to English. Tokenization was thus performed for these instances as well, as seen in Table 1. This particular phase is finished to reduce the level of complexity associated with labeling. Furthermore, since the chosen tokenization is predicated on Universal Dependence (UD), determining the primary and auxiliary verbs is crucial. As a result, the primary verbs in the phrase are recognized, and the other verb tenses are made auxiliary. This allows the relationship to be determined based on auxiliary while developing the tree bank parser. As a result, auxiliaries form the foundation for both the dependency parser and POS labeling. Let us take the line "இன்று இரவு நான் போகவில்லை" as an example. In this sentence, the word "போகவில்லை" assumes the role of the main verb, while the remaining words become the auxiliary verbs. After that, this auxiliary verb is subjected to POS tagging, which is described in the next section.

Table 1. Sample list of words and their corresponding affixes used in tokenization.

Tokenization affixes	Tokenization example words
Clitics	Un, um, unadhu, avadhu
Auxiliary verbs	Villai, ulladhu,ullar,ullanar, pathathu,mudiyaadhu etc.,
Postpositions	Andru, netru, indru, thavira etc.,
Postpositions	Ap, ic etc.,
Particles	Aha and their variants etc.,

- **Part Of Speech mapping (POS)**

The morphological tagging of the data is carried out using the transformer model's POS tagging. The BeRT-based tagger was able to identify five different entities: person, time, location, date, and organization. Thus, the positional tagging method, as mentioned in [28], was utilized to perform the tagging. This tagging wasn't done manually; however, the TnT tagger [20] was used. Figure 4 shows the tag format used here.

Figure 4 considers the Tamil word "பாடினார்" and shows the output of the tagging done for the same. As shown, there was consideration of nine entities for each word, and their corresponding tagging is done. The tagging was distinct as the morphological characteristics of the tokens were considered for the tagging, and the main tags like tense, nouns, and pronouns were correctly assigned, and the accuracy was verified manually. Table 2 provides the tagging details and the list for the first two entities POS and SubPOS as they are tagged using the Roberta model.

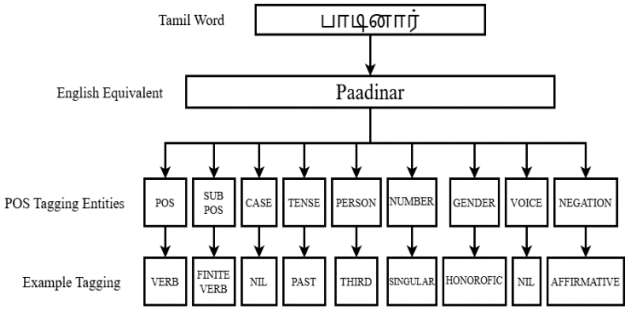


Figure 4. POS Tagging entities considered with the combination of BeRT and TnT Tagger.

Table 2. Categories of the POS tagging.

Tag	Description	Example
Case	Accusative Dative Genitive Instrumental Locative Nominative	Sociative AI Kku Al In Il Ntu Out
Tense	Past Future Present Tenseless	Ar Um Ikkum Illai
Person	1st person 2nd person 3rd person	En Al Ikkum
Number	Plural singular	Kal Aar
Gender	Feminine Masculine Neuter Honorific Animate Inanimate	Val Nin Tatu Avar Yar unused
Voice	Active Passive	Tatu ratu
Negation	Affirmative Negation	Parra Yattu

- **Dependency Relation Mapping**

This is where the different terms' dependencies are constructed and annotated. This mostly attempts to identify the root word and establish how it relates to the child's words. Thus, by joining the dependent words to those of the root word, a structure resembling a tree is created based on the morpho-syntactic analysis. You may think of this structure as a knowledge graph with a root and an edge. A label describing the relationship between the parent and child nodes is stored on each edge. Here, we adhere to the 39 dependency connections as acl, acl:relcl, advcl, advcl:cond, advmod, advmod:emph, amod, aux, aux:neg, aux:pass, case, cc,

compound, compound:lvc, compound:redup, conj, det, fixed, iobj, mark, nmod, nmod:poss, nsubj, nsubj:nc, nsubj:pass, nummod, obj, obl, obl:agent, obl:arg, obl:cmpr, obl:inst, obl:lmod, obl:pmod, obl:tmod, punct, root, vocative, xcomp that make up the dependency relation analytical functions that are discussed in [6]. Some of the sample dependency relations formed using the rule-based parser mentioned in [29] are shown in Figures 5 and 6.

Figure 5 represents the root universal dependency that is formed by pairing the root word to that of the verb, noun, pronoun, proper noun, adjective, and number. Figure 6 forms the relationship aux: neg, punct, nsubj. Random sentences are given as input and their corresponding universal dependency is obtained from the rule-based parser and they are shown in Figures 5 and 6. The same mapping rules are used in the transformer model, however, the morpho-syntactic analysis was done with a different tagger. These dependency parsing results are used in performing a comparison to that of the transformer-generated universal dependency after obtaining the outputs from the Roberta model. This comparison is done to observe the changes that are obtained with the different tagging methods and the manual annotation. It is been observed that the results of the transformer matched in the universal dependency parser and thus the morphosyntactic analysis done using the transformer network was further used for the conversion process to signs.

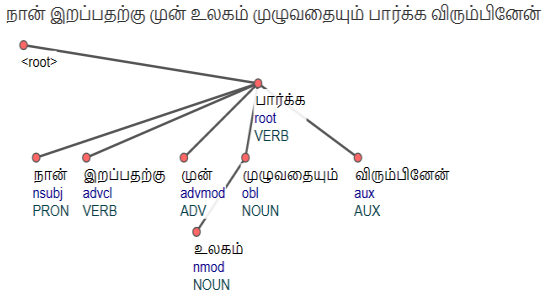


Figure 5. Root relation universal dependency.

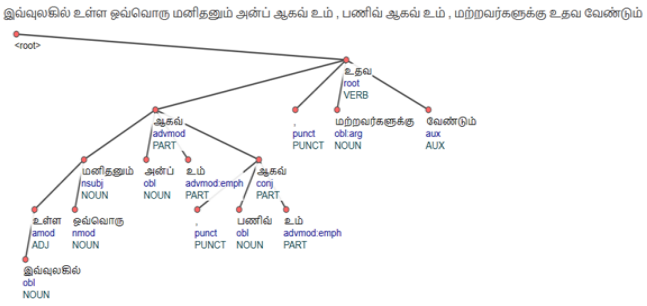


Figure 6. Punct relation universal dependency.

Figure 5 represents the root universal dependency that is formed by pairing the root word to that of the verb, noun, pronoun, proper noun, adjective, and number. Figure 6 forms the relationship aux: neg, punct, nsubj. Random sentences are given as input and their corresponding universal dependency is obtained from the rule-based parser and they are shown in Figures 4 and 5. The same mapping rules are used in the transformer model, however, the morpho-syntactic analysis was done with a different tagger. These dependency parsing results are used in performing a comparison to that of the transformer-generated universal dependency after obtaining the outputs from the Roberta model. This comparison is done to observe the changes that are obtained with the different tagging methods and the manual annotation. It is been observed that the results of the transformer matched in the universal dependency parser and thus the morphosyntactic analysis done using the transformer network was further used for the conversion process to signs.

3.2. Stage 2-Sign Language Production Network (SLPN)

The sign language production performs the process of converting the gloss to its corresponding poses. Thus, our SLPN inputs the gloss probabilities obtained from the RMSA, which are then used to solve the motion graph for the generation of poses for text. This pose is conditioned to produce the sign sequences with the encoder and decoder networks for the production of signs for the given text. In this section, let us discuss each of the components that frame this SLPN.

3.2.1. Text -2-Pose

The neural machine translation is carried out using the RMSA that performed the syntactic analysis of the same and the gloss sequence is thus obtained based on the morpho-syntactic structure of the sentence.

Given a spoken Tamil language sentence S_n={w_1, w_2, w_3… … … .. w_n} as input, the transformer encoder mapped the sequence to a latent representation as in Equation (1).

$$output_{w_1:w_n}, hidden_n^e = Robencoder(S_n) \quad (1)$$

Where the $output_{w_1:w_n}$ is the encoder output for each of the words and the $hidden_n^e$ shows the hidden form of the sentence and here in this case it denotes the morpho-syntactic analysis representation. Using the hidden and encoded values for each word the probability of gloss is generated. Here the mapping is sequence mapping that considers the encoded and hidden representation. This encoder output from the transformer network when it gets passed on to the decoder generates the probability distribution over glosses as shown in Equation (2)

$$prob(gloss_t) = Robdecoder(gloss_{t-1}, hidden_{t-1}^d) \quad (2)$$

Where $gloss_t$ represents the gloss generated at time step t and $hidden_{t-1}^d$ denotes the previous decoder's hidden state.

Specifically, a transformer-based encoder and decoder are used because of the inbuilt attention

mechanism that helps the decoder obtain enough information about the previously hidden representation and also avoids the long-term dependency problem.

3.2.2. Motion Graph

The next step in the process of SLP is the pose generation for the glosses generated by the transformer network. To achieve this, we rely on a motion graph through which skeletal poses are generated for the different gloss formations. A motion graph is a directed graph that consists of nodes and edges defined as where M denotes the motion graph, N represents the nodes and they are the various movements in a sequence, and thus $N \in n_i$

where $n=\{1, 2, 3 \ldots i\}$ and thus n is the number of motion primitives in a sequence. Since the gloss generated here is the sign for each of the individual words, the gloss boundary is the stopping criteria for cutting the pose.

Here every node does not mean a single pose, a single node can have more than one pose because the decoder generates a gloss at every time step *t* and thus all the glosses that are generated for that time step *t* could be represented as a single node. Thus, the motion graph is simply a combination of the different gloss information recorded at time step *t*, and thus the node is a combination of the current motion primitive and the previous one, and that is represented using the probability distribution so, the node of the motion graph is motion (m_i) and their probabilities $p(m_i)$.

The edges in the motion graph denote the transition from one node to another. Here this transition is the change of motion primitives. So typically an edge comprises two points denoted as $E \in edge_{i,j}$ where $edge_{i,j}$ is the connection between the two nodes *i* and *j*. Every edge formation includes the motion blending information and the probability distribution.

Having defined the node and edges next is to understand the way the nodes and the edges get the data. Thus, the number of letters in a sequence frames the number of nodes and the edge information is obtained from the morpho-syntactic knowledge graph. So here the edges get populated directly by the transformer decoder network based on the prior probability value as shown in Equation (3).

$$edge_{t-1,t} = p(x_t|x_{t-1}) \tag{3}$$

Equation (3) defines the motion transition information that makes the edge move from one node to another using the knowledge graph information and the corresponding probability function. Thus, Equation (3) can be modified as mentioned in Equation (4) including the decoder information.

$$edge_{t-1,t} = Robdecoder(gloss_{t-1}, hidden_{t-1}^{d}) \tag{4}$$

For the given sentence, the sequence for the motion graph starts with the beginning node and at each motion step, the list of hypotheses from the knowledge graph is considered. At every edge formation, the hypothesis is expanded to include the new motion primitive as shown in Equation (5)

$$hyp_m^t = hyp_m^{t-1}, x_t^* \tag{5}$$

where a set of motions of *m* at time step *t* and the x_t^* denotes the starting node that is chosen based on the probability distribution computation for edge as mentioned in Equation (6).

$$x_t^* = argmax_x p(x_t|x_{t-1}) \tag{6}$$

Using Equation (6) the hypothesis is expanded till that specific motion sequence ends.

3.3. Pose2Video

This process of generating high-quality videos for the sign poses obtained using the motion graph is performed using the Generative Adversarial Network (GAN). GAN is generally preferred for this task due to its structure. GAN comprises two networks, termed generator and discriminator, that get trained together as the newer samples generated by the generator are verified for their truthfulness by the discriminator. So, the generator aims to maximize the discriminator's likelihood of false prediction, and the discriminator aims to minimize the same with the correct identification. This min-max strategy can thus produce samples that are very close to the real ones.

Thus, a simple GAN network is used for the creation of the video sequences and the generator is fed with the pose label and the root pose. The generator attempts to condition the images based on the previous poses so that the temporal elements are rightly captured. The image generated can be termed as $G_{im}=(pose_{label}, pose_{root})$ is formulated with the input being passed on to the generator encoder which is up convoluted by the decoder using the pose map information.

4. Experimentation and Results

This work aims to build the entire pipeline for performing the translation from a Tamil sentence to its corresponding sign video. So before evaluating the components of the pipeline quantitatively and qualitatively, first of all, get a detailed idea about the dataset and the preprocessing adopted for the construction of the gloss and pose. After this explanation, the results are shown in the form of translations of Tamil sentences to gloss sequences, pose sequences, and video sequences.

4.1. Datasets

The process of generating sign language to the spoken text requires a large-scale dataset in the form of text and their corresponding sign language sequences. One such larger source is the RWTH-PHOENIX-Weather 2014T which comprises the sign interpretations of the weather.

Since such a dataset is not available for the Tamil language, so here for this work 100 Tamil sentences and their corresponding glosses are generated.

The dataset comprises 100 sign language video sequences performed by a single signer in a controlled environment. The video sequences range in size from 10 MB to 25 MB, with each video lasting no longer than 2 minutes. The chosen sentences for the videos represent commonly used phrases such as "Where are you going?" and "I am coming with you." The videos are captured from various angles to comprehensively capture the signer's movements, with a particular focus on hand and facial expressions, which are essential for conveying information through signs. Each video frame undergoes auto skin segmentation and normalization to ensure consistent brightness throughout the sequence. Semantic word variants are grouped and assigned gloss labels based on the Chaii Dataset. Furthermore, linguistic annotations include gloss labels, onset and offset times for each sign, hand movement intervals, and morphological classifications, all meticulously conducted using the Roboflow package in Python. Ultimately, the sequences are labeled with sign glosses and their corresponding spoken language translations, enhancing the accessibility and comprehensibility of the dataset.

Along with this data, the transformer network is also trained with the PHOENIX14T data and the SMILE sign language assessment dataset. This training is done since the architecture involves the translation of sentences to English and thus certain signs can be derived from this dataset as well. This ensures that there is no alignment between the text and sign language sequences. Having a wide variety of data ensures that knowledge transfer between different datasets can happen and thus the prediction capability of the system becomes good. This is also being done to understand whether training with multiple datasets helps in translation between different spoken and sign languages.

4.2. Data Preprocessing

The sign language videos were trimmed and for each of the clips manual labeling of the word was carried out and the corresponding character mapping to the words and the corresponding sign glosses were obtained using the ASL dataset. The landmarks in the form of CSV were converted to JSON format upon which the preprocessing for Morpho syntactic analysis is carried out as explained in section 3.1.

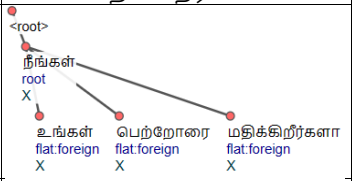
4.3. Tamil Sentence to Gloss

The SLPN network as mentioned prior is an encoder and decoder network that possess 6 layers with 500 gated recurrent units. The transformer attention mechanism is the default one utilized for the context vector generation. The optimizer here is Adam with a learning rate of 〖10^(-3) for 50 epochs. The dropout probability was 0.2 while training. The text-to-pose generation reported an average time of 0.5s per gloss translation using the AMD RYZEN 9000 processor with 8GB cache.

4.3.1. Translating Tamil to ASL Gloss

As described in section 3.1., gloss generation here happens after the morpho analysis of the Tamil sentence. To verify the correctness of the model, a specific video recorded for the sentence "நீங்கள் உங்கள் பெற்றோரை மதிக்கிறீர்களா?" is considered and all the intermediate visual results shown in this paper are for this specific sentence. Figure 7 presents the actual video sequence for the sentence and its corresponding word-by-word sign translation using ASL annotation. This is done to show the matching of certain symbols that were performed by a native Tamil language speaker. As mentioned, this step is crucial because the motion graph is guided by this annotation and the knowledge graph that preserves the structure of the sentence. Table 3 presents the neural machine translation carried out for some of the sentences.

Table 3. Neural machine translation using the RMSA network.

Type	Sentence and Gloss
Tamil sentence	நீங்கள் உங்கள் பெற்றோரை மதிக்கிறீர்களா
Morpho analysis	
English equivalent for the actual sentence	Are you respecting your parents
ASL gloss	RESPECTING PARENTS
Text2gloss	YOU RESPECTING PARENTS

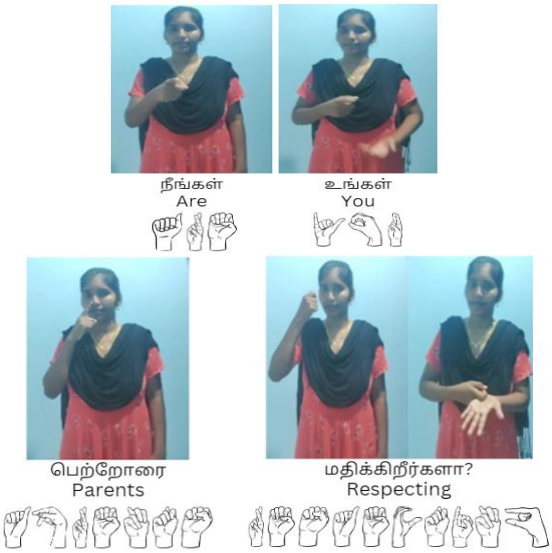


Figure 7. Gloss annotation using ASL alphabet set and the matching symbols with the annotation gloss and actual word representation in Tamil language.

Our experimentation about the text2gloss is to understand whether the transformer-based morpho syntactic analyzed gloss generated is equivalent to the ASL gloss. From Figure 4 and Table 1, it is clear that the proposed RMSA network can produce gloss for Tamil sentences that match the ASL gloss. This matching is observed in terms of the characters as well as the words. Though the matching is not exact, the gloss generated was close to the meaning of the sentence. Now this gloss was used for the generation of the human pose maps.

The quantitative analysis of the translation ability of the text2pose method was done using the BLEU, ROGUE, and Word Error Rate (WER) metrics. These metrics remain the standard evaluation metrics with machine translation. Since this is a first-of-a-kind work considering the Tamil language, state of the art text2gloss method is unavailable for making a comparison. So here we compare our results to Text2Sign [22] and this we have carried out just to make a comparison with other language glosses. Our results are shown in Table 4.

Table 4. Quantitative measures in terms of BLEU, ROGUE, and WER.

Approach	Training						Testing					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	WER	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	WER
Text2Sign [22]	50.15	32.47	22.30	16.34	48.42	4.83	50.67	32.35	21.54	15.26	48.10	4.53
Proposed	51.23	30.12	24.21	14.32	52.12	2.54	55.26	35.26	17.21	13.32	54.10	6.28

As observed from Table 2, Text2Sign has achieved a high BLEU-4, and the performance of Tamil Lang TSP with smaller gloss is good in terms of the BLEU scores and ROUGE. Though we have considered the grammar for the sentences, the temporal grammar associated with sign languages is something that is not considered by the text and since this can become more with a larger gloss there is an expected decrease with the quantitative values.

4.3.2. Text-2-Pose

The qualitative evaluation for the translation of Tamil sentences to human pose sequence with the solution obtained from the motion graph is presented in Figures 8 and 9. In the first case, we show the key frames that are obtained for the actual sentence and the second case represents the key frames extracted for the gloss. The relevancy of the keyframes in both cases are identical and hence it is evident that the conditioning of poses for a specific gloss has happened.

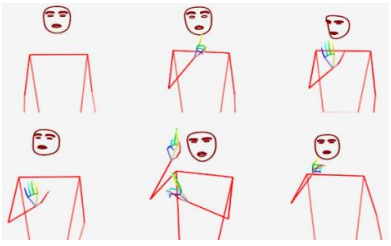


Figure 8. The pose sequence obtained for the actual sentence "நீங்கள் உங்கள் பெற்றோரை மதிக்கிறீர்xகளா."

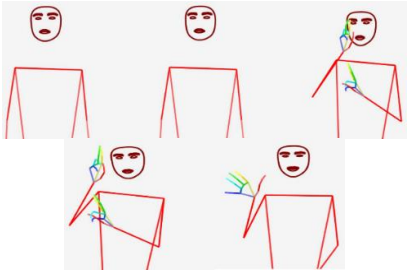


Figure 9. The pose sequence obtained for the gloss "பெற்றோரை மதிக்கிறீர்களா உங்கள்."

5. Conclusions

Through this work, the first spoken language to sign language translation for the Tamil language is presented. Instead of using complex avatar-based approaches, this method uses the motion graph to learn the morph-syntactic analysis of the Tamil sentence that combines the neural machine translation and graph through which the pose sequences are made. Then the same pose sequences are used to condition the GAN network for the production of sign video.

The morph-syntactic analysis of the sentence that was carried out as the neural machine translation network is used for obtaining the pose sequences from the motion graph. These results are analyzed quantitatively and qualitatively and also we have utilized the GAN network for the production of synthetic images that map the pose sequences to them and a photo-realistic sign generation is made.

The initial work done for Tamil language sentences is pioneering, but there is still significant room for improvement. Currently, the dataset is limited to just 100 sentences, which hinders real-time application of the system. This underscores the need for further training with a larger dataset comprising more sentences. In addition, integrating facial expressions alongside hand movements in sign language can enhance communication. Furthermore, the inclusion of more nonmanual signs is on our radar for future work. Our present approach falls short in terms of definition when compared to avatar-based methods, so our ongoing efforts focus on mapping the motion graph to achieve higher definition in our future work.

Acknowledgment

The Authors express their heartfelt thanks to Dr. R.I.Minu, Professor at SRM Institute of Science and Technology, Chennai, for their constant support, Encouragement, and help in completing this research work.

Author Contributions

Conceptualization, S.T., and R.I.M.; Methodology, S.T., and R.I.M.; Software, S.T.; Validation, S.T. and R.I.M.; Formal Analysis, S.T.; Investigation, S.T. and R.I.M.; Resources, S.T. and R.I.M.; Data Curation, S.T.; Writing-original draft preparation, S.T.; Writing-review and editing, S.T.; Visualization, S.T.; Supervision, R.I.M.

References

[1] Aliwy A. and Ahmed A., "Development of Arabic Sign Language Dictionary Using 3D Avatar Technologies," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 1, pp. 609-616, 2021. https://ijeecs.iaescore.com/index.php/IJEECS/article/view/22518/14493

[2] Balaha M., El-Kady S., Balaha H., Salama M., Emad E., Hassan M., and Saafan M., "A Vision-Based Deep Learning Approach for Independent-Users Arabic Sign Language Interpretation," *Multimedia Tools and Applications*, vol. 82, no. 5, pp. 6807-6826, 2023. https://link.springer.com/article/10.1007/s11042-022-13423-9

[3] Basiri S., Taheri A., Meghdari A., Boroushaki M., and Alemi M., "Dynamic Iranian Sign Language Recognition Using an Optimized Deep Neural Network: An Implementation via a Robotic-Based Architecture," *International Journal of Social Robotics*, vol. 15, no. 4, pp. 599-619, 2023. https://link.springer.com/article/10.1007/s12369-021-00819-0

[4] Bora J., Dehingia S., Boruah A., Chetia A., and Gogoi D., "Real-Time Assamese Sign Language Recognition Using Media Pipe and Deep Learning," *Procedia Computer Science*, vol. 218, pp. 1384-1393, 2023. https://doi.org/10.1016/j.procs.2023.01.117

[5] Bowden R., "Learning to Recognize Dynamic Visual Content from Broadcast Footage," *Engineering and Physical Sciences Research Council*. https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/I011811/1

[6] Brants T., "TnT-A Statistical Part-of-Speech Tagger," *in Proceedings of the 6th Applied Natural Language Processing*, Seattle, pp. 224-231, 2000. https://aclanthology.org/A00-1031/

[7] Brour M. and Benabbou A., "ATLASLang NMT: Arabic Text Language into Arabic Sign Language Neural Machine Translation," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 9, pp. 1121-1131, 2021. https://doi.org/10.1016/j.jksuci.2019.07.006

[8] Cox S., Lincoln M., Tryggvason J., Nakisa M., Wells M., Tutt M., and Abbott S., "Tessa, a System to Aid Communication with Deaf People," *in Proceedings of the 5th international ACM Conference on Assistive Technologies*, Edinburgh, pp. 205-212, 2002. https://dl.acm.org/doi/10.1145/638249.638287

[9] Das Chakladar D., Kumar P., Mandal S., Roy P., Iwamura M., and Kim B., "3D Avatar Approach for Continuous Sign Movement Using Speech/Text," *Applied Sciences*, vol. 11, no. 8, pp. 3439, 2021. https://doi.org/10.3390/app11083439

[10] Das S., Biswas S., and Purkayastha B., "A Deep Sign Language Recognition System for Indian Sign Language," *Neural Computing and Applications*, vol. 35, no. 2, pp. 1469-1481, 2023. https://link.springer.com/article/10.1007/s00521-022-07840-y

[11] Das S., Imtiaz M., Neom N., Siddique N., and Wang H., "A Hybrid Approach for Bangla Sign Language Recognition Using Deep Transfer Learning Model with Random Forest Classifier," *Expert Systems with Applications*, vol. 213, pp. 118914, 2023. https://doi.org/10.1016/j.eswa.2022.118914

[12] Duarte A., Palaskar S., Ventura L., Ghadiyaram D., DeHaan K., Metze F., Torres J., and Giro-i-Nieto X., "How2Sign: A Large-Scale Multimodal Dataset for Continuous American Sign Language," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, pp. 2735-2744, 2021. https://ieeexplore.ieee.org/document/9577749

[13] Ebling S., Camgoz N., Braem P., Tissi K., Sidler-Miserez S., Stoll S., Hadfield S., Haug T., Bowden R., Tornay S., Razavi M., and Magimai-Doss M., "SMILE Swiss German Sign Language Dataset," *in Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, pp. 4221-4229, 2018. https://aclanthology.org/L18-1666/

[14] Efthimiou E., Fotinea S., Hanke T., Glauert J., Bowden R., Braffort A., Collet C., Maragos P., and Lefebvre-Albaret F., "The Dicta-Sign Wiki: Enabling Web Communication for the Deaf," *in Proceedings of the Computers Helping People with Special Needs: 13th International Conference*, Linz, pp. 205-212, 2012. https://link.springer.com/chapter/10.1007/978-3-642-31534-3_32

[15] Elliott R., Glauert J., Kennaway J., and Marshall I., "The Development of Language Processing Support for the ViSiCAST Project," *in Proceedings of the 4th International ACM Conference on Assistive Technologies*, Arlington, pp. 101-108, 2000. https://dl.acm.org/doi/10.1145/354324.354349

[16] Farooq U., Mohd Rahim M., and Abid A., "A Multi-Stack RNN-Based Neural Machine Translation Model for English to Pakistan Sign Language Translation," *Neural Computing and Applications*, vol. 35, no. 18, pp. 13225-13238, 2023. https://link.springer.com/article/10.1007/s00521-023-08424-0

[17] Filhol M. and McDonald J., "The Synthesis of Complex Shape Deployments in Sign Language," *in Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages*, Marseille, pp. 61-68, 2020. https://www.sign-lang.uni-hamburg.de/lrec/pub/20.html

[18] Forster J., Schmidt C., Koller O., Bellgardt M., and Ney H., "Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather," *in Proceedings of the 19th International Conference on Language Resources and Evaluation*, Reykjavik, pp. 1911-1916, 2014. https://aclanthology.org/L14-1472/

[19] Gibet S., Lefebvre-Albaret F., Hamon L., Brun R., and Turki A., "Interactive Editing in French Sign Language Dedicated to Virtual Signers: Requirements and Challenges," *Universal Access in the Information Society*, vol. 15, no. 4, pp. 525-539, 2016. https://link.springer.com/article/10.1007/s10209-015-0411-6

[20] Hajic J., *Disambiguation of Rich Inflection: Computational Morphology of Czech*, Karolinum Press, Charles University, 2004. https://doi.org/10.2478/jazcas-2019-0067

[21] Kar P., Reddy M., Mukherjee A., and Raina A., "INGIT: Limited Domain Formulaic Translation from Hindi Strings to Indian Sign Language," *ICON*, vol. 52, pp. 53-54, 2007. https://www.cse.iitk.ac.in/users/purushot/papers/ingit.pdf

[22] Kothadiya D., Bhatt C., Saba T., Rehman A., and Bahaj S., "SIGNFORMER: DeepVision Transformer for Sign Language Recognition," *IEEE Access*, vol. 11, pp. 4730-4739, 2023. https://ieeexplore.ieee.org/document/10011551

[23] Kumar Attar R., Goyal V., and Goyal L., "State of the Art of Automation in Sign Language: A Systematic Review," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 4, pp. 1-80, 2023. https://dl.acm.org/doi/abs/10.1145/3564769

[24] Manzano D., "English to ASL Translator for Speech2Signs," 2021. https://www.semanticscholar.org/paper/ENGLISH-TO-ASL-TRANSLATOR-FOR-SPEECH2SIGNS-Manzano/88d6a9573b9a2d0cf54c90e947f4fbe12578fbca

[25] McDonald J., Wolfe R., Schnepp J., Hochgesang J., Jamrozik D., Stumbo M., Berke L., Bialek M., and Thomas F., "An Automated Technique for Real-Time Production of Lifelike Animations of American Sign Language," *Universal Access in the Information Society*, vol. 15, no. 4, pp. 551-566, 2016. https://link.springer.com/article/10.1007/s10209-015-0407-2

[26] Mittal A., Kumar P., Roy P., Balasubramanian R., and Chaudhuri B., "A Modified LSTM model for Continuous Sign Language Recognition Using Leap Motion," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 7056-7063, 2019. https://ieeexplore.ieee.org/document/8684245

[27] Othman A. and Jemni M., "English-ASL Gloss Parallel Corpus 2012: ASLG-PC12," *in Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, Istanbul, pp. 151-154, 2012. https://www.sign-lang.uni-hamburg.de/lrec/pub/12019.html

[28] Ramasamy L. and Zabokrtsky Z., "Prague Dependency Style Treebank for Tamil," *in Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, pp. 1888-1894, 2012. https://aclanthology.org/L12-1242/

[29] Ramasamy L. and Zabokrtsky Z., "Tamil Dependency Treebank," *in Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*, Tokyo, pp. 82-94, 2011. https://ufal.mff.cuni.cz/~ramasamy/tamiltb/1.0/html/

[30] Saunders B., Camgoz N., and Bowden R., "Everybody Signs Now: Translating Spoken Language to Photo-Realistic Sign Language Video," *arXiv Preprint*, vol. arXiv:2011.09846, pp. 1-11, 2020.

[31] Saunders B., Camgoz N., and Bowden R., "Progressive Transformers for End-To-End Sign Language Production," *in Proceedings of the Computer Vision-ECCV, 16th European Conference*, Glasgow, pp. 687-705, 2020. https://link.springer.com/chapter/10.1007/978-3-030-58621-8_40

[32] Shin H., Kim W., and Jang K., "Korean Sign Language Recognition Based on Image and Convolution Neural Network," *in Proceedings of the 2nd International Conference on Image and Graphics Processing*, Singapore, pp. 52-55, 2019. https://dl.acm.org/doi/10.1145/3313950.3313967

[33] Shin J., Miah A., Hasan M., Hirooka K., Suzuki K., Lee H., and Jang S., "Korean Sign Language Recognition Using Transformer-Based Deep Neural Network," *Applied Sciences*, vol. 13, no. 5,

pp. 3029, 2023. https://doi.org/10.3390/app13053029

[34] Stoll S., Camgoz N., Hadfield S., and Bowden R., "Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks," *in Proceedings of the 29th British Machine Vision Conference British Machine Vision Association*, Newcastle, pp. 1-12, 2018. https://www.researchgate.net/publication/326811903_Sign_Language_Production_using_Neural_Machine_Translation_and_Generative_Adversarial_Networks

[35] Stoll S., Camgoz N., Hadfield S., and Bowden R., "Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks," *International Journal of Computer Vision*, vol. 128, no. 4, pp. 891-908, 2020. https://link.springer.com/article/10.1007/s11263-019-01281-2

[36] Tokuda M. and Okumura M., *Assistive Technology and Artificial Intelligence: Applications in Robotics, User Interfaces, and Natural Language Processing*, Springer, 2006. https://link.springer.com/chapter/10.1007/bfb0055973

[37] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A., Kaiser L., and Polosukhin I., "Attention is all you Need," *arXiv Preprint*, vol. arXiv:1706.03762v2, pp. 1-16, 2017. https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[38] Veale T. and Conway A., "Cross-Modal Comprehension in ZARDOZ an English to Sign-Language Translation System," *in Proceedings of the 7th International Workshop on Natural Language Generation*, Maine, pp. 249-252, 1994. https://dl.acm.org/doi/10.5555/1641417.1641450

[39] Ventura L., Duarte A., and Giro-i-Nieto X., "Can Everybody Sign Now? Exploring Sign Language Video Generation from 2D Poses," *arXiv Preprint*, arXiv:2012.10941, pp. 1-4, 2020. https://arxiv.org/pdf/2012.10941

[40] World Health Organization, Deafness and Hearing Loss, https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, Last Visited, 2024.

[41] Xie P., Cui Z., Du Y., Zhao M., Cui J., Wang B., and Hu X., "Multi-Scale Local-Temporal Similarity Fusion for Continuous Sign Language Recognition," *Pattern Recognition*, vol. 136, pp. 109233, 2023. https://doi.org/10.1016/j.patcog.2022.109233

[42] Zelinka J. and Kanis J., "Neural Sign Language Synthesis: Words are our Glosses," *in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Snowmass, pp. 3395-3403. 2020. https://ieeexplore.ieee.org/document/9093516

[43] Zhao L., Kipper K., Schuler W., Vogler C., Badler N., and Palmer M., "A Machine Translation System from English to American Sign Language," *in Proceedings of the Envisioning Machine Translation in the Information Future: 4th Conference of the Association for Machine Translation in the Americas*, Cuernavaca, Mexico, pp. 54-67, 2000. https://link.springer.com/chapter/10.1007/3-540-39965-8_6

[44] Zwitserlood I., Verlinden M., Ros J., Van der Schoot S., and Netherlands T., "Synthetic Signing for the Deaf: Esign," *in Proceedings of the Conference and Workshop on Assistive Technologies for Vision and Hearing Impairment*, 2004. https://www.semanticscholar.org/paper/SYNTHETIC-SIGNING-FOR-THE-DEAF-%3A-eSIGN-Zwitserlood-Verlinden/df8bdfaabf2e043e22cf0c8b544b619f33d96e05

Thillai Sivakavi S is currently pursuing Ph.D. in the Department of Computing Technologies at the Faculty of Engineering and Technology, Kattankulathur Campus, Chennai, India.

Minu R I serves as a Professor in the Department of Computing Technologies at SRM Institute of Science and Technology, Kattankulathur. She holds a B.E degree in Electronic and Communication Engineering from Bharadhidasan University, an ME degree in Computer Science Engineering from Anna University, and a Ph.D. in Computer Science Engineering from Anna University, focusing on "Ontology Enhanced Image Retrieval for Semantic Web." She had completed a Post Doc in the year 2023 at University of Louisiana at Lafayette, USA focusing on "Quantum Deep Learning algorithm for Image Recognition." With 16 years of teaching experience. She has authored three books and more than 50 research papers in refereed international journals and conferences. Currently, she boasts 749 citations on Google Scholar, with an h-index of 15 and an i10 index of 22. Her research interests encompass a wide array of fields including Computer Vision, Artificial Intelligence, Ontology Learning, Machine Learning, Internet of Things, Smart Grid, Edge Computing, and Mixed Reality.