www.ijarp.com ISSN 2456-9992 Page: 01-21

International Journal Advanced Research Publications

Review Article

Volume: 02
Issue: 08

DIGITAL MARGINALIZATION OF LOW-RESOURCE LANGUAGES: EVALUATING GENERATIVE AI TRAINING DATA GAPS FOR SRI LANKAN SIGN LANGUAGE (SSL)

***S. Earnest Bosco, V.G.S Pradeepika, D.M.Y.B Dissanayake**

Department of Information Technology, Sri Lanka Institute of Advanced Technological Education (SLIATE), Sri Lanka.

Article Received: 01 July 2026, Article Revised: 21 July 2026, Published on: 08 August 2026

***Corresponding Author: S. Earnest Bosco**

Department of Information Technology, Sri Lanka Institute of Advanced Technological Education (SLIATE), Sri Lanka.

Doi: https://doi-doi.org/101555/ijarp.3489

1. ABSTRACT

Generative artificial intelligence (GenAI), which includes text-to-image, text-to-video and synthetic 3D avatar technologies, has the potential to improve digital accessibility. However, the large datasets used to train these multimodal models carry serious geographic, linguistic and cultural biases, and they support mostly high-resource languages such as American Sign Language (ASL). This paper examines the lack of training data for Sri Lankan Sign Language (SSL), a unique visual-spatial language used by the Deaf community in Sri Lanka. The study is a structured literature review. It examines the published record on generative AI and sign language, the documented state of SSL datasets and recognition research, the literature on digital marginalization and algorithmic content distribution, and the literature on data sovereignty and community governance. The findings show that SSL is absent from the documented training data of generative models, that the country's existing resources are small, recognition oriented and unevenly documented, and that the technical gap is inseparable from a governance gap. In addition, the literature indicates that algorithm-based platforms reduce the reach of SSL content because it does not carry spoken audio in Sinhala, Tamil or English. The paper proposes a community-governed data sovereignty framework suited to the legal and technological context of Sri Lanka, in order to prevent the algorithmic erasure of SSL.

KEYWORDS: generative AI; Sri Lankan Sign Language; low-resource languages; digital marginalization; data sovereignty; sign language datasets; assistive technology.

2. INTRODUCTION

2.1. Background

Generative artificial intelligence (GenAI) is no longer confined to research laboratories. In the past few years, tools that can create images, videos, speech and even three-dimensional avatars from a text prompt have entered everyday use, and many of them are now marketed as accessibility solutions. Text-to-image systems can help a person with low vision to understand a photograph. Text-to-video systems can turn written information into animated content. Synthetic avatars can present news bulletins, read out announcements, or act as virtual sign language interpreters. For the Deaf community in Sri Lanka, this seems like an important moment. A technology that can turn spoken language into sign language, or sign language into text, could remove barriers that have existed for decades.

This promise, however, depends on the data that these models are trained on. GenAI models learn from very large collections of text, images and video, and most of this material is taken from the internet. The internet is not a fair sample of the world. It is dominated by a small number of large languages and by the cultures of wealthy countries. A model that has seen millions of hours of American Sign Language (ASL) video, and very little of any other sign language, will serve ASL users well and serve everyone else poorly. This gap is not accidental. It follows from decisions about which data to collect, which languages to fund, and which markets to build for.

2.2. Problem statement

Language technology research uses the term "low-resource language" for a language that has very little usable training data. Sign languages are among the most under-resourced languages of all. ASL, British Sign Language, Chinese Sign Language and a few others enjoy large public datasets and commercial interest. The rest of the world's sign languages have almost none. Sri Lankan Sign Language (SSL) belongs to this second group.

SSL is a complete and independent language. It is visual and spatial in nature, with its own grammar, its own handshapes and its own non-manual markers, such as facial expression, head movement and eye gaze, which carry grammatical meaning. It is not a signed version of Sinhala or Tamil. It is the native language of the Sri Lankan Deaf community, used across the country, and it differs clearly from the sign languages of neighbouring countries. Any

generative system that wishes to serve this community must learn these features correctly. Whether it can do so is exactly the question this review examines.

Local research interest in SSL has grown steadily in recent years. Abhishek (2024) released SSL400, a dataset built at the Informatics Institute of Technology, which has become a reference collection for the language. Liyanaarachchi and colleagues (2021) published a signing dataset for Sinhala Sign Language through the Sri Lanka Technological Campus. Munasinghe and colleagues (2023) explored a wearable sensor based approach for interpretation. Silva and Perera (2021) combined a convolutional neural network with Scale Invariant Feature Transform to translate SSL into Sinhala text, and researchers at the South Eastern University of Sri Lanka (2025) reported a real-time static SSL recognition system using EfficientNet-B0. This body of work is valuable, but it is almost entirely concerned with recognition, that is, turning signs into text. It says very little about the generative side of artificial intelligence, the side that produces video, images and avatars. That is where the marginalization this review describes actually takes place.

2.3. Digital marginalization and the local creator economy

We use the term "digital marginalization" to describe a situation in which a community is systematically left out of digital technologies, not because its members refuse to use them, but because the technologies were never built with that community in mind. For a language community, the exclusion appears in a simple form: the models do not know the language, so the models cannot work for it. This exclusion operates at two levels. At the level of the model itself, a generative system that has never been trained on SSL video will make serious visual errors when asked to produce SSL, misreading local non-manual markers and regional handshapes. At the level of the platform, the reach of authentic SSL content is limited because social media algorithms are tuned to content that has spoken audio in Sinhala, Tamil or English, and signed content usually has none. Together these two mechanisms reduce both the quality and the visibility of SSL in the digital space.

The second level deserves emphasis because it touches the people who create content. Deaf creators in Sri Lanka who record and publish SSL content are part of a small local creator economy. When their content is not recommended, not translated and not indexed, they receive fewer viewers, fewer opportunities and less income than hearing creators. The problem is therefore not only technical. It is also economic and cultural, and any meaningful solution must treat it as such.

2.4. Aim and objectives

The aim of this review is to examine, on the basis of the published literature, the training data gaps that exclude Sri Lankan Sign Language from generative AI, and to propose a way of addressing them. The review was carried out with four specific objectives:

1. To survey the literature on GenAI and sign language, and to establish which sign languages are represented in the documented training data of generative systems.
2. To review the published accounts of SSL datasets and SSL recognition research in Sri Lanka, in order to describe what data resources exist and how they are oriented.
3. To examine the literature on digital marginalization, algorithmic content distribution and the creator economy, in order to understand how low-resource languages are disadvantaged beyond the model itself.
4. To draw on the literature on data sovereignty and community governance in order to propose a framework suited to the Sri Lankan context.

2.5. Research questions

Four questions guided the review:

1. What does the published literature say about which sign languages are represented in generative AI training data, and is SSL among them?
2. What is the documented state of SSL datasets and SSL research in Sri Lanka, and how does it compare with high-resource sign languages?
3. What does the literature say about how platforms and algorithmic systems marginalize low-resource and signed content?
4. What does the literature say about community-governed data sovereignty, and how could it apply to SSL in Sri Lanka?

2.6. Method in brief

This is a literature review. It draws on peer-reviewed papers, conference proceedings, dataset documentation and policy documents. The method is described fully in Section 3. The review does not collect new data. It synthesizes what is already published, and on that basis it develops an argument about digital marginalization and a framework for addressing it.

2.7. Significance of the review

This review is significant for a number of reasons. First, it brings together two bodies of literature that are usually kept apart: the technical literature on sign language AI and the

social literature on digital marginalization. Second, it directs attention to the generative side of AI, which has received almost no attention in Sri Lankan sign language research. Third, it treats the Deaf community not simply as a source of data, but as a community with the right to decide what happens to its language. Fourth, it proposes a community-governed data sovereignty framework that is suited to the legal and technological environment of Sri Lanka.

2.8. Organization of the paper

The remainder of this paper is arranged as follows. Section 2 reviews the relevant literature. Section 3 describes the method of the review. Section 4 presents the findings from the literature. Section 5 discusses what these findings mean for the Sri Lankan context. Section 6 proposes the community-governed data sovereignty framework. Section 7 concludes the paper and suggests directions for future work.

3. LITERATURE REVIEW

3.1. Generative AI and digital accessibility

The rapid spread of generative AI has produced a large and enthusiastic literature about its potential for people with disabilities. Researchers have argued that text-to-speech, text-to-image and text-to-video systems can make information more accessible, that synthetic avatars can take over repetitive interpreting tasks, and that automated captioning can open up video content (Bragg et al., 2019; Duarte et al., 2021). There is truth in these claims. The same generation of models that can write an essay or draw a picture can also describe a photograph for a person with low vision, or produce a spoken narration of a document for someone who cannot read it.

What this literature tends to leave out is the question of whom the models are built for. Accessibility features are added to products that already exist and already serve a profitable market. The accessibility of the marginal user is treated as an add-on. For the Deaf community this has a particular consequence. Sign languages are not small dialects of spoken languages. They are separate languages with separate grammars, and a hearing company that builds an accessibility feature rarely has the training data or the linguistic knowledge to support a sign language properly. The result is that the accessibility promise of GenAI is real for some languages and almost entirely absent for others.

3.2. Training data and language bias in multimodal models

A large body of evidence shows that the datasets behind multimodal models are geographically and linguistically skewed. Studies of image-text collections repeatedly find that the majority of content originates from North America and Western Europe, and that the languages represented are a handful of large ones (Bragg et al., 2023). This bias is not a neutral property of the data. It becomes a property of the model. A model trained on such data produces images, captions and videos that reflect the visual world of its training sources, and it performs worse, often much worse, for people, places and languages that are underrepresented.

The consequences for sign languages are especially severe, because sign languages are expensive to collect and to annotate. A usable dataset needs video, it needs signers who are fluent, and it needs linguistic annotation, which in turn needs trained deaf or hearing linguists. The high cost of good data is the main reason that ASL, British Sign Language and Chinese Sign Language have large corpora while nearly all other sign languages have almost none. What the wider literature describes as the data divide in AI is, for sign languages, more like a data chasm.

3.3. Sign language datasets and the high-resource divide

The existing sign language datasets tell a clear story. On one side are the benchmark corpora of the well-resourced languages, such as How2Sign for ASL and the RWTH-PHOENIX corpus for German Sign Language, each with thousands of hours of annotated continuous signing (Camgoz et al., 2020; Duarte et al., 2021). These datasets have made possible the current wave of recognition and translation systems, including large generative models for sign language production. On the other side are the low-resource sign languages, for which the largest available collection may hold a few hundred videos of isolated signs. Surveys of the field consistently find that a very small share of all publicly available sign language data belongs to languages outside the high-resource group, and that many sign languages have no AI-related dataset at all (Bragg et al., 2019; Bragg et al., 2023).

This imbalance matters for more than fairness. It changes what researchers can study. Because the big datasets are in ASL and a few European and East Asian languages, the methods that dominate the literature, transformer-based recognition, sign production models, and synthetic avatar systems, are designed and tested for those languages. Whether these

methods transfer to a language like SSL, with its own handshapes and non-manual markers, is an open question, and the scarcity of data makes it a difficult question to answer.

3.4. Sri Lankan sign language research

Within Sri Lanka, research on SSL has grown steadily but remains concentrated on recognition. Abhishek (2024) introduced SSL400, a dataset of Sri Lankan Sign Language collected at the Informatics Institute of Technology, and this collection is now widely used as a benchmark for the language. Liyanaarachchi and colleagues (2021) developed a signing dataset for Sinhala Sign Language at the Sri Lanka Technological Campus, adding further resources for training and testing. On the technical side, Munasinghe and colleagues (2023) proposed a wearable sensor based approach to interpreting SSL, while Silva and Perera (2021) combined a convolutional neural network with Scale Invariant Feature Transform to translate SSL into Sinhala text. Most recently, researchers at the South Eastern University of Sri Lanka (2025) reported a real-time static sign recognition system built on EfficientNet-B0, and reported strong accuracy on their test set.

Taken together, this work demonstrates that SSL recognition is feasible. What is striking is what the work does not cover. Almost all of it targets isolated signs or the static alphabet, evaluates on data collected by the researchers themselves, and stops at the point where a recognised sign becomes text or a spoken word. There is no published system that generates SSL, no work that studies how a generative model would handle the non-manual markers of the language, and no study that examines whether commercial GenAI tools can serve Sri Lankan Deaf users at all. In other words, the literature has built the recognition side of the pipeline and has left the generation side, which is the side where the marginalization described in this paper occurs, essentially untouched.

3.5. Digital marginalization and the creator economy

The concept of digital marginalization has been used in development and media studies to describe how groups can be pushed to the edges of digital life even when they are physically connected to the internet. Two mechanisms are commonly identified. The first is representational: the technologies do not know the group's language or culture, so the group's needs are invisible in product design. The second is algorithmic: the platforms that carry content are tuned to optimise for engagement, and engagement tends to favour content with spoken audio in a major language. Signed content, which is visual and has no audio track in

Sinhala, Tamil or English, is poorly matched to such systems, so it is recommended less often and reaches fewer people.

This second mechanism has a direct economic effect, and it is relevant to Sri Lanka because it operates on the local creator economy. Deaf creators who publish SSL content depend on the same platforms as hearing creators for reach and income. When the algorithm suppresses their content, the marginalization is experienced as a loss of viewers, opportunities and livelihood, not merely as an inconvenience. The literature on the global creator economy has documented this pattern for smaller spoken languages, but almost no attention has been given to sign languages, and none to SSL.

3.6. Data sovereignty and community governance

Faced with these conditions, several fields have turned to the idea of data sovereignty, the principle that a community should have control over the collection, use and distribution of data about itself. In sign language research this argument has been made most strongly in connection with indigenous sign languages and with the Deaf communities of the Global South. The argument runs as follows. Language data is not a free resource to be extracted. It is an expression of a community's identity, and its collection and use carry risks, including the risk that a model trained on community data will be deployed in ways the community did not agree to. For this reason, researchers argue that datasets should be governed by the communities they come from, through consent, benefit sharing and community oversight.

Sri Lanka provides a specific legal and technological context for such an approach. The country has data protection legislation that is still being implemented, an active civil society around disability rights, and a research community that is small but connected. No framework yet exists that applies the principle of data sovereignty to SSL specifically. This paper takes the position that such a framework is needed, and that it must be designed with the Deaf community as the governing partner, not as a source of data for others to exploit.

3.7. Summary of gaps

The review of the literature reveals three clear gaps. First, the training data gap itself: SSL has small datasets built for recognition, and no representation in the corpora that feed generative models. Second, a research gap: local work has not examined whether generative systems can handle SSL at all, and the question of non-manual markers and regional handshapes is unstudied. Third, a governance gap: the question of who should own and control SSL data has not been asked in the Sri Lankan context, even though the principle of

community governance is well established elsewhere. These three gaps define the space in which this study operates, and they are taken up in turn in the sections that follow.

4. METHODOLOGY

4.1. Type of review

This is a structured literature review, sometimes called a scoping review. It is suited to a topic like this one because the aim is to map what is known across several separate bodies of work, rather than to measure the effect of a single intervention. The review follows the spirit of the PRISMA-ScR guidance, which sets out how scoping reviews should report their search, screening and synthesis in a transparent way (Tricco et al., 2018). A full registration was not undertaken, but the search strategy, the inclusion criteria and the data extraction steps are set out below so that the review can be repeated.

The review covers four areas of the literature: (a) generative AI and sign language, (b) SSL datasets and recognition research in Sri Lanka, (c) digital marginalization, algorithms and the creator economy, and (d) data sovereignty and community governance of language data.

4.2. Sources searched

The search was conducted across the main scholarly databases used in this field: IEEE Xplore, ACM Digital Library, Scopus, Google Scholar and OpenAlex. These sources were chosen because sign language AI research is published mostly in computer science venues, which are well covered by IEEE, ACM and Scopus, while Google Scholar and OpenAlex help to capture smaller conference papers and theses that the subscription databases may miss. In addition to the scholarly databases, the review examined the documentation of major publicly available generative models and the published description of sign language datasets, since dataset documentation is a form of published literature in its own right.

4.3. Search strategy

Search strings were built around four blocks of terms, matching the four areas of the review: The searches were run between February and May 2026. Where a database allowed it, the search was limited to titles, abstracts and keywords, and to publications in English. No lower date limit was applied, so that early work on SSL recognition could be captured.

4.4. Inclusion and exclusion criteria

A source was included in the review if it met all of the following conditions:

(a) it was a peer-reviewed paper, a conference proceeding, a thesis, a dataset paper or a documented model card;

(b) it concerned sign language AI, a Sri Lankan or South Asian sign language, digital marginalization, or data sovereignty as these relate to language; and

(c) it was available in English.

Table 1: Structure of search blocks and example search terms.

Block	Focus	Example terms
A	GenAI and sign language	"Sign language" generation, "sign language" avatars, "generative AI" accessibility, text-to-video sign language
B	SSL research and datasets	"Sri Lankan Sign Language", "Sinhala Sign Language", SSL400, SSL50
C	Marginalization and platforms	"Digital marginalization", "low-resource language" AI, sign language social media reach, algorithmic content
D	Data sovereignty	"Language data sovereignty", "indigenous data sovereignty", community governance sign language

A source was excluded if it was a news article without a research component, a product brochure, or a publication that mentioned sign language only in passing without contributing evidence. Grey literature was not excluded outright, because for a topic as small as SSL, official reports and dataset documentation carry evidence that peer-reviewed papers do not, but such sources were treated with more caution and their claims were checked against other sources where possible.

4.5. Screening and data extraction

The records retrieved from the searches were first deduplicated and then screened by title and abstract. Where the title and abstract were not enough to judge relevance, the full text was consulted. After screening, the included records were read in full and the following items were extracted into a structured form: the year and venue of publication, the area of the review it addressed, the sign language or languages concerned, the dataset or model described if any, and the main claim or finding.

For the SSL research block, a separate extraction was made of the reported characteristics of each dataset or system, namely the number of videos or samples, the number of sign classes,

the type of data, and whether the resource was publicly available. These figures are reported in Section 4.2, and where a paper did not report a figure, it is recorded as not reported.

3.6 Synthesis

The included literature was synthesized narratively, area by area. A narrative synthesis is appropriate here because the four bodies of literature use different methods and different measures, so a statistical combination of results is neither possible nor meaningful. The findings from the four areas were then brought together into an overall account of how SSL is marginalized in the generative AI space, and this account forms the basis for the framework in Section 6.

4.6. Limitations of the method

A review of this kind has limitations, and they should be stated openly. First, the topic is small and fast moving, and the generative model ecosystem changes more quickly than the academic literature can track. Some of the most recent models may be described only in preprints or in developer documentation, which are covered but are not always stable sources. Second, the review is limited to English-language material. Most of the relevant literature is in English, but a Tamil-language or Sinhala-language source might exist that this review has not captured. Third, the published literature on SSL is thin, so the review depends heavily on a small number of papers and on dataset documentation. These limitations define the scope of the findings that follow.

6. FINDINGS

6.1. What the literature says about GenAI and sign language coverage

The literature on generative AI and sign language is consistent in one respect: it concerns almost entirely the high-resource sign languages. The benchmark systems that have driven the field, the transformers that perform sign language recognition and translation, and the generative models that produce signed output are all built and evaluated on ASL, German Sign Language, Chinese Sign Language or British Sign Language (Camgoz et al., 2020; Duarte et al., 2021; Walsh et al., 2024). Surveys of the field report that the majority of all publicly available sign language data belongs to a small handful of languages, and that the documentation of what exists is often incomplete (Bragg et al., 2019; Bragg et al., 2023).

The literature does not report any generative system, dataset or model with documented support for Sri Lankan Sign Language. Nothing in the published record suggests that SSL appears in the training data of any major generative model. The documented picture is the

same for South Asian sign languages as a group: they appear in the recognition literature, where researchers in India, Pakistan, Bangladesh and Sri Lanka have built systems for their own languages, but they do not appear in the generative literature. The conclusion that follows from the published sources is straightforward. In the generative space, SSL is absent.

6.2. What the literature says about SSL datasets and research

The published record on SSL research in Sri Lanka is small but real. It is dominated by recognition systems built on self-collected data. The main documented collections are summarized in Table 1.

Table 2: Existing Sri Lankan Sign Language (SSL) datasets and systems.

Dataset or system	Year	Institution	Reported content	Public
SSL50	2023	University of Moratuwa	1,110 videos, 50 sign classes, 6 signers	Yes
SSL400	2024	Informatics Institute of Technology	reference collection for SSL	Yes
Sinhala SL signing dataset	2021	Sri Lanka Technological Campus	signing data for Sinhala SL	not stated
Wearable sensor system	2023	IEEE publication	sensor based interpretation data	no
CNN and SIFT system	2021	ICARC publication	self-collected static signs	no
EfficientNet-B0 system	2025	South Eastern University of Sri Lanka	self-collected static signs	no

Sources: Abhishek (2024); Charuka et al. (2023); Liyanaarachchi et al. (2021); Munasinghe et al. (2023); Silva and Perera (2021); South Eastern University of Sri Lanka (2025).

Three patterns in this literature matter for the present review. First, the collections are small, and they are unevenly documented. Some papers report the number of videos, classes and signers; others report only accuracy, and at least one reports nothing about the data itself. Second, the work is oriented almost entirely toward isolated signs and the static alphabet.

Continuous signing, which is what real communication consists of and what a generative model must handle, is not represented. Third, the resources were built for recognition, that is, for classifying a sign into a category, not for generation, which requires paired signing and meaning at scale. The literature contains no account of a Sri Lankan dataset built for the generative task, and no account of a generative SSL system at all.

When this record is compared with what the literature reports for high-resource languages, the gap is of a different order. How2Sign alone provides over eighty hours of continuous ASL video with full annotation (Duarte et al., 2021), and the corpora used to train sign language transformers are measured in thousands of hours (Camgoz et al., 2020). The Sri Lankan literature describes collections of a few thousand samples of isolated signs. The comparison is not between two sizes of the same thing. It is between two different kinds of resource, and only one of those kinds can feed a generative model.

6.3. What the literature says about digital marginalization and platforms

The literature on digital marginalization provides the frame for interpreting the technical findings above. Several strands are relevant. First, the bias literature shows that web-crawled datasets and the models trained on them are skewed toward a small number of languages and toward the content of wealthy countries, and that this skew becomes a property of the models themselves (Bragg et al., 2023). Second, the disability and accessibility literature documents that Deaf users face barriers in digital spaces that are not solved by connectivity alone, and that language is a central part of those barriers (Wedasinghe and Wicramaarchchi, 2014; Udugama et al., 2024). Third, the media and platform literature describes how algorithmic content distribution rewards content that platforms can classify, and signed content, which has no audio track in a major language, is hard for platforms to classify and therefore reaches fewer people.

Taken together, these strands describe the two-level marginalization set out in the introduction. The model cannot produce SSL correctly because it has never seen SSL, and the platform will not distribute SSL effectively because the content does not fit the audio based logic of recommendation. The literature supports both claims, and it supports the further point that these two levels reinforce each other. Content that is not seen does not attract the attention, investment or research that would improve the models.

6.4. What the literature says about data sovereignty and community governance

The fourth body of literature concerns who should own and control language data. In sign language research this argument has been made most strongly in connection with indigenous sign languages and with the Deaf communities of the Global South. The argument runs as follows. Language data is not a free resource to be extracted. It is an expression of a community's identity, and its collection and use carry risks, including the risk that a model trained on community data will be deployed in ways the community did not agree to (Bragg et al., 2023; Zeshan et al., 2013). For this reason, researchers argue that datasets should be governed by the communities they come from, through consent, benefit sharing and community oversight.

For Sri Lanka specifically, the literature adds important context. Udugama and colleagues (2024) report that SSL is not yet a fully standardised language, that regional variation is high, that the interpreter workforce is very small, and that deaf and hard of hearing Sri Lankans are marginalized and under-supported. These conditions bear directly on data governance. A community that is not confident in the status of its own language, and that has few interpreters to gloss and validate data, needs a governance arrangement that builds trust and capacity rather than assuming them. The framework proposed in Section 6 is designed with these conditions in mind.

6.5. Summary of the findings

The literature, read across the four areas, yields three consistent findings. First, SSL is absent from the documented training data of generative AI. Second, the Sri Lankan resources that exist are small, recognition oriented and unevenly documented, and they do not provide the paired, annotated, continuous data that generation requires. Third, the literature on marginalization and on data sovereignty shows that the technical gap is inseparable from a governance gap, and that any remedy must place the Deaf community at the centre of data ownership. These findings are the basis for the discussion and the framework that follow.

7. DISCUSSION

7.1. What the literature shows as a whole

Read as a whole, the literature describes a form of digital marginalization that operates at two levels. The first level is the model. Generative systems have no documented SSL training data, and the published Sri Lankan record explains why: the country's collections are small, oriented to isolated signs, unevenly documented, and none is built for generation. A model

trained on web-crawled data dominated by high-resource sign languages will produce visually wrong SSL, particularly in the non-manual markers and regional handshapes that carry the grammar of the language. The literature offers no reason to expect otherwise, and no published account of any attempt to close this gap on the generative side.

The second level is the platform. The literature on algorithmic distribution shows that signed content, which has no spoken audio track in a major language, is hard for platforms to classify and therefore reaches fewer people. This is not a minor effect. It is a ceiling on the reach and income of Deaf creators, and it means that even authentic SSL content is marginalized in the same digital space where generative models are unable to produce it. Between the two levels, the language is disadvantaged in both directions: it cannot be produced well, and what is produced is not seen.

7.2. Why the gap is structural rather than incidental

The SSL data gap is often described as a resource problem, as if more money and more recording effort would close it. The literature reviewed here suggests that the gap is structural, in three ways.

First, the gap is a matter of task orientation. The entire documented Sri Lankan dataset infrastructure was built for recognition. Generation requires something different: paired signing and meaning, continuous signing rather than isolated signs, and linguistic annotation. Expanding the existing recognition datasets will not produce this. A different kind of data collection is required.

Second, the gap is a matter of governance. The data sovereignty literature shows that communities are wary of data collection when they have not been consulted about it, and this wariness is rational in a context where a community has been studied without benefit to itself (Bragg et al., 2023). If the Deaf community does not trust the process, it will not contribute signing, and without community contribution no generative dataset can be built. The technical gap and the trust gap are therefore the same gap.

Third, the gap is a matter of capacity beyond data. The country has very few qualified interpreters, and SSL is still being standardized (Udugama et al., 2024). Any dataset effort depends on people who can gloss, translate and validate signing, and those people are scarce. A framework for SSL data must therefore be realistic about the resources the country actually has.

7.3. Comparison with other contexts

The situation of SSL is not unique, but it is not exactly the same as that of other low-resource sign languages. The literature reports the same concentration of resources in a handful of languages worldwide (Bragg et al., 2023), and the same pattern of prototype systems that never reach deployment. Where Sri Lanka differs is in the relative strength of its existing data efforts. The country has real datasets, SSL400 and SSL50, which is more than many low-resource sign languages have, but these are recognition resources built by institutions, with no generative purpose and no documented community governance.

The comparison with the international data sovereignty literature is also instructive. Researchers working with indigenous and low-resource sign languages have argued that data collection must be governed by the community, with consent and benefit sharing at the centre (Zeshan et al., 2013; Bragg et al., 2023). The principle is well established internationally. What the Sri Lankan literature adds is the specific local context in which the principle would have to be applied: a language still being standardized, a very small interpreter workforce, and a research community that is small but connected.

7.4. Implications for policy and practice

Three practical implications follow from the literature. First, for researchers, the review suggests that building a generative SSL resource is a legitimate and urgent research target, but it cannot be built the way recognition datasets were built. It must be built as a community resource, with consent, with linguists, and with the generative task in mind from the start.

Second, for institutions, the review suggests that dataset work should be coordinated rather than scattered. The existing collections come from different institutions with different formats and different documentation standards. A coordinated effort built on shared standards would make the existing work far more valuable than it is in isolation.

Third, for policymakers, the review connects the data gap to the wider policy context. The Sign Language Bill is close to being enacted, and the country has committed to disability inclusion in principle. A community-governed data framework for SSL would give practical meaning to these commitments, and it would prevent a repeat of the pattern documented in the literature, where a community's language is collected and studied without the community's control.

7.5. Limitations of the review

The limitations of this review should be acknowledged alongside its findings. The review is limited to English-language material, so a source in Sinhala or Tamil might have been missed. The published literature on SSL is thin, and the findings depend heavily on a small number of papers and on dataset documentation, some of which is incomplete. The generative model ecosystem changes quickly, and the most recent developments may appear only in preprints or developer documentation rather than in peer-reviewed literature. These limitations do not invalidate the findings. They define their scope, and they point to the need for continued monitoring of a fast-moving field.

8. A Community-Governed Data Sovereignty Framework for SSL

8.1. Principles

The findings of this review point to a single conclusion: the future of SSL in generative AI depends on data that the Deaf community itself controls. This section proposes a framework for that control. It draws on the data sovereignty literature reviewed in Section 2.6, and it is adapted to the specific conditions of Sri Lanka, namely a language still being standardized, a very small interpreter workforce, and a small but connected research community.

The framework rests on five principles. First, community primacy. The Deaf community is the owner of SSL data, not a source of it. Every decision about collection, use and sharing is taken with the community, and the community holds the final say. Second, free, prior and informed consent. No data is collected without consent that is given in SSL itself, not in a written form the signer cannot read. Third, benefit sharing. The benefits of any dataset or model, whether research credit, income, or accessible tools, flow back to the community that contributed the data. Fourth, transparency. The community always knows what data exists, who holds it, and how it is used. Fifth, capacity building. The framework must train Deaf signers and linguists to gloss, validate and govern their own data, so that the community is not dependent on outside experts forever.

8.2. Components

The framework has four components.

The first component is a governance body. A national SSL data governance council, drawn from Deaf community organisations, the few qualified interpreters in the country, researchers and representatives of the relevant ministries, would hold authority over SSL data. The council would approve or reject data collection proposals, set the terms of consent, and audit

how data is used. Its membership would be majority Deaf, so that the community governs itself rather than being advised by others.

The second component is a consent and licensing protocol. Every recording would be collected under a documented licence agreed with each signer, and the licence would set out what the data may be used for, whether it may be used for commercial purposes, and whether it may be used to train generative models at all. The protocol would be written in plain Sinhala and Tamil and presented in SSL, so that consent is real rather than nominal. This component answers the trust concern identified in the literature: a community that has been studied without consent will not contribute data unless the consent process is designed for it.

The third component is the data infrastructure itself. The framework calls for a single national SSL repository, built on open standards, that records every dataset, its size, its signers, its licensing and its documentation. The repository would bring together the scattered collections described in Section 4.2, the SSL400 and SSL50 datasets and the others, under common documentation rules. It would be governed by the council, not by any single institution. The repository would begin as a recognition resource, but it would be designed so that generative resources, paired signing and meaning at scale, could be added as the community's capacity grows.

The fourth component is capacity building and linkage to policy. The framework would fund the training of Deaf annotators and linguists, support the glossing and validation work that the country currently lacks the people to do, and connect the repository to the policy process. The Sign Language Bill, when enacted, would provide a legal anchor for community ownership of the language. The repository and the council would give that ownership a working form.

8.3. How the framework addresses the findings

Each finding of this review maps to a part of the framework. The absence of SSL from generative training data is addressed by the repository, which would create the kind of resource that generative models need. The recognition orientation of the existing data is addressed by the design decision to build for generation from the start. The governance gap is addressed by the council and the consent protocol, which place the community at the centre. The capacity gap is addressed by the training component. And the platform marginalization described in Section 4.3, while not directly fixable by a data framework, is made visible by the repository and the council, which give the community a stronger base from which to

demand platform treatment. The framework does not claim to solve every part of the problem. It claims to create the conditions under which the problem can be solved on the community's own terms.

8.4. Feasibility in the Sri Lankan context

A framework of this kind is ambitious, but it is not unrealistic, and it should be judged against the alternative. The alternative is that SSL data continues to be collected by individual research groups, each with its own format and its own purposes, with no community control and no national record of what exists. The framework asks for less, not more, than that in some ways: one repository instead of many private collections, one consent standard instead of none, and one body that can speak for the data. The scarce resources are already there in small measure, the institutions, the researchers and the interpreters. What is missing is coordination and community authority. The framework supplies both.

The main risks are three. The first is that the council could become a body that speaks for the community without being accountable to it, and the design answers this by making the membership majority Deaf and by requiring audits. The second is that the repository could become a storehouse that serves researchers without benefiting the community, and the design answers this by the benefit sharing principle and by the capacity building component. The third is that the framework could be ignored by the large companies that build generative models, and the design answers this, imperfectly, by giving the community a clear licence that companies must respect if they wish to use the data. No framework can compel a large company to act. But a clear licence and a single governing body are a far stronger position than the current one, where nothing is documented and nothing is governed.

9. CONCLUSION

This paper examined, through a structured review of the published literature, the training data gaps that exclude Sri Lankan Sign Language from generative AI. The review found that SSL is absent from the documented training data of generative models, that the country's existing resources are small, oriented to recognition, and unevenly documented, and that the technical gap is inseparable from a governance gap. It also found that the literature describes a two-level form of digital marginalization: generative models cannot produce SSL correctly because they have never seen it, and platforms will not distribute SSL effectively because signed content does not fit the audio-based logic of recommendation.

The review makes three contributions. First, it brings together the technical literature on sign language AI and the social literature on digital marginalization, which are usually kept apart, and shows that they describe the same problem. Second, it directs attention to the generative side of AI, which has received almost no attention in Sri Lankan sign language research. Third, it proposes a community-governed data sovereignty framework, built on community primacy, free and informed consent, benefit sharing, transparency and capacity building, that is suited to the legal and technological context of Sri Lanka.

The framework proposes a national SSL data governance council, a consent and licensing protocol presented in SSL itself, a single national repository built on open standards, and a capacity building programme linked to policy. Its purpose is to prevent the algorithmic erasure of SSL before generative systems become widely adopted in Sri Lanka, rather than to correct the damage after the fact.

Several directions for future work follow from this review. Empirical research is needed to test how commercial generative models actually perform on SSL tasks, since the present review can only report what the documentation and the literature reveal. Linguistic work is needed to support the standardization of SSL and to document its non-manual markers and regional variation, which are precisely the features that generative models fail on. And participatory work with the Deaf community is needed to design the governance council and the consent protocol in detail, with the community itself, rather than on its behalf. None of this work can be done without the community, and none of it should be. SSL belongs to the Deaf community of Sri Lanka, and its future in the digital space belongs to them as well.

10. REFERENCES

1. Abhishek, Y. (2024). *SSL400 - Sri Lankan Sign Language Dataset*. Informatics Institute of Technology (IIT), Sri Lanka.
2. Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., Caselli, N., Huenerfauth, M., Kacorri, H., Verhoef, T., Vogler, C., & Ringel Morris, M. (2019). Sign language recognition, generation, and translation: An interdisciplinary perspective. *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2019)*.
3. Bragg, D., Caselli, N., Hochgesang, J. A., Huenerfauth, M., Katz-Hernandez, L., Koller, O., Kushalnagar, R., Lillo-Martin, D., Menzies, L., Occhino, C., Seita, M., Shaw, A., Smith, A., Cates, D., & Vogler, C. (2023). The FATE landscape of sign language AI

datasets: An interdisciplinary perspective. *ACM Transactions on Accessible Computing*, 16(3).

4. Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*.
5. Charuka, T., et al. (2023). SSL50: Sinhala sign language dataset. University of Moratuwa, Sri Lanka.
6. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., & Giro-i-Nieto, X. (2021). How2Sign: A large-scale multimodal dataset for continuous American Sign Language. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021)*.
7. Liyanaarachchi, K. L. P. P., Shakya, D., Herath, T., Vithanage, N., & Udugama, L. S. K. (2021). Signing dataset for the Sinhala Sign Language. Sri Lanka Technological Campus (SLTC), Sri Lanka.
8. Munasinghe, N., Jayalal, S., & Wijayasiriwardhane, T. (2023). Interpretation of Sri Lankan Sign Language: A wearable sensor-based approach. *IEEE Conference Proceedings*.
9. Silva, R., & Perera, C. (2021). Sri Lankan Sign Language to Sinhala text using convolutional neural network combined with Scale Invariant Feature Transform (SIFT). *International Conference on Advanced Research in Computing (ICARC)*.
10. South Eastern University of Sri Lanka. (2025). Real-time Sri Lankan static sign language system using EfficientNet-B0. *Sri Lanka Journal of Technology*.
11. Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., et al. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467-473.
12. Udugama, L. S. K., et al. (2024). Sign language usage of deaf or hard of hearing Sri Lankans. *Journal of Deaf Studies and Deaf Education*, 29(2), 187-198.
13. Wedasinghe, W., & Wicramaarchchi, W. (2014). ICT adoption and usage among disabled persons in Sri Lanka. *IEEE International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*.
14. Walsh, H., et al. (2024). SignGPT: Generative pre-training for sign language. *arXiv preprint*.
15. Zeshan, U., et al. (2013). *Sign Languages in Village Communities*. De Gruyter Mouton / Ishara Press.