https://doi.org/10.24303/lakdoi.2025.33.2.115

Challenges in Deep Learning-Based Analysis of Korean Sign Language: Through the Lens of American Sign Language Research

Yong-hun Lee
(Chungnam National University)

Lee, Yong-hun. (2025). Challenges in deep learning-based analysis of Korean sign language: Through the lens of American sign language research. *The Linguistic Association of Korea Journal, 33*(2), 115-135. Sign language is a fully developed linguistic system using visual-gestural elements such as hand movements, facial expressions, and spatial organization. While deep learning has advanced American Sign Language (ASL) research, applying these methods to Korean Sign Language (KSL) faces challenges due to KSL's classifier predicates, spatial referencing, and topic-comment structures. This paper critically reviews ASL-based deep learning in Sign Language Recognition (SLR), Sign Language Production (SLP), and Sign Language Translation (SLT) to assess their adaptation for KSL. In this review, SLR covers the automatic recognition of sign sequences from visual input, SLP addresses the generation of natural sign gestures from text or speech, and SLT focuses on translating between sign and spoken languages. Methodologically, we conduct a comparative literature review of state-of-the-art deep learning models, analyzing their architectures, training strategies, and evaluation metrics within each subfield (SLR, SLP, SLT). We examine linguistic differences between ASL and KSL, noting difficulties in gesture synthesis, spatial modeling, and non-manual feature integration. We highlight limitations of direct ASL-to-KSL model transfer and propose multi-modal learning, expanded datasets, and enhanced spatial encoding to advance KSL processing technologies.

Key Words: American sign language, Korean sign language, recognition, production, translation

1. Introduction

Sign language is a fully developed linguistic system that enables communication through visual-gestural means, such as hand movements, facial expressions, and body posture. Unlike spoken languages, sign languages are structured in a spatial-temporal manner, incorporating grammar which is unique to their modality. Each sign language, such as Korean Sign Language (KSL) and American Sign Language (ASL), contains an independent system and is not merely a manual representation of spoken language. Many linguists recognize sign languages as natural languages, because they exhibit complex linguistic structures. Their studies may provide insights into Cognitive Linguistics, language acquisition, and the diversity of human communication systems.

Research on sign language is crucial for improving accessibility and fostering inclusivity for deaf and hard-of-hearing individuals. It enables the development of technologies such as sign language recognition, generation, and translation, that enhance communication between deaf and hearing communities. Additionally, sign language research contributes to linguistic theories by examining how visual languages function in comparison to spoken ones. It also supports language preservation efforts, ensuring that diverse sign languages are properly documented and rarely standardized. With advancements in artificial intelligence and deep learning, sign language research is becoming increasingly relevant for developing assistive tools that promote communication and education.

The goal of this study is to critically review the research papers on ASL and KSL, focusing on studies that implement and analyze ASL using deep learning techniques. By examining these studies, the research aims to explore challenges that may arise when applying similar computational methods in ASL to KSL. While ASL has been extensively studied in deep learning-based sign language recognition, KSL presents unique linguistic features that may pose difficulties in implementation. This study will identify potential challenges specific to KSL and highlight key considerations for future research and technological advancements in KSL processing.

The literature for this study was selected through a systematic search of research papers available on the arXiv repository (https://www.arxiv.org/), focusing on works directly related to SLR, SLP, and SLT. From the initial collection of papers, papers were categorized according to their primary focus on the datasets used or the models implemented, enabling a structured analysis of methodological trends. Within each

category, additional filtering was applied based on scholarly impact, with citation counts and community engagement serving as key indicators of relevance and influence. By combining topical relevance, methodological diversity, and citation-based significance, the resulting corpus provides a balanced foundation for critical review and comparative analysis.

This paper is organized as follows. Section 2 discusses the characteristics of ASL and KSL respectively, highlighting their similarities and differences. Section 3 covers sign language recognition (SLR), focusing on deep learning approaches in ASL. Section 4 explores sign language production (SLP), examining how natural and grammatically accurate signs can be produced by AI techniques. Section 5 delves into sign language translation (SLT), addressing methods for converting sign language into text or speech. Section 6 presents discussions on the findings and challenges which are identified in the previous three sections. Section 7 concludes the study by summarizing key insights and proposing directions for future research in KSL deep learning applications.

2. ASL vs. KSL

2.1. Characteristics of ASL[1)]

American Sign Language (ASL) is a fully developed visual-gestural language used by Deaf communities in the United States of America and parts of Canada. It exhibits complex linguistic structures and sociolinguistic variations similar to spoken languages; but it differs in its modality, because it relies on hand movements, facial expressions, and body posture. ASL is not merely a representation of English but an independent language system with its own grammar and syntax. It serves as a crucial means of communication for the Deaf community and plays a significant role in their culture. Studying ASL enhances understanding of language diversity and the cognitive processes behind visual communication.

ASL originated from the American School for the Deaf (ASD), founded in 1817 in Hartford, Connecticut. Laurent Clerc, a Deaf educator from France, brought elements of French Sign Language (LSF), which combined with indigenous sign systems like Martha's Vineyard Sign Language to form ASL. Over time, ASL evolved through interactions in

1) The summary in this section is primarily based on Lucas & Bayley (2010).

Deaf communities, education, and social networks. Its history reflects the resilience and adaptability of Deaf culture. Studying ASL's historical development provides insight into language evolution and highlights the importance of preserving and documenting sign languages for future generations.

ASL consists of distinct phonological components, including handshape, movement, location, palm orientation, and nonmanual markers such as facial expressions and body posture. These elements function similarly to phonemes in spoken languages. Variations also occur within these components, as signers may modify handshapes or orientations without significantly changing meaning. Nonmanual markers also add grammatical information, such as indicating questions or emotions. As spoken languages have dialectal and phonetic differences, ASL displays variation influenced by region and social factors. Understanding the phonological aspects is essential for sign language recognition technology and linguistic research.

ASL exhibits lexical variation, with multiple signs existing for the same concept, much like synonyms in spoken languages. For example, words like *birthday*, *picnic*, and *Halloween* can have different signs depending on regional dialects, social groups, or historical influences. This variation reflects cultural diversity within the Deaf community and is influenced by factors such as age and ethnicity. Just as English speakers use different words for the same object depending on region or background, ASL users may prefer different signs based on their linguistic environment. Recognizing lexical variation is vital for effective ASL education and translation technologies.

ASL follows a topic-comment structure, where the topic is established first, followed by a description or action. It is also a *pro*-drop language, meaning that subject pronouns can be omitted when context makes them unnecessary. For example, *I feel* can be expressed simply as *feel* without any ambiguity. ASL also employs verb agreement, particularly with directional verbs that incorporate subject and object references through movement. For instance, the sign *help* can move from the signer toward another person to mean *I help you* or in the opposite direction to mean *You help me*. These features demonstrate ASL's grammatical complexity.

ASL, like spoken languages, exhibits sociolinguistic variation influenced by region, age, gender, and ethnicity. Different regions of the USA have distinct sign variations, forming ASL dialects. Older signers may retain signs no longer widely used by younger generations. Gender differences have also been observed, with variations in signing styles. Additionally, African American signers historically developed a distinct ASL dialect due

to segregation in Deaf education. Educational background and exposure to Deaf culture further influence sign language variation. Understanding these sociolinguistic factors is crucial for ensuring inclusive sign language education, interpretation, and research.

ASL's visual-gestural modality results in unique linguistic properties that are distinct from spoken languages. Unlike spoken languages which are linear, ASL can convey multiple pieces of information simultaneously. Facial expressions can indicate a question while hands sign a sentence. Additionally, ASL uses spatial grammar, where signers reference physical space for grammatical relationships. Some signs can be produced with one or both hands, depending on context. These modality-driven features make ASL efficient and expressive but also present challenges in computational recognition. Understanding ASL's modality is essential for improving sign language processing and translation technologies.

Residential Deaf schools played a crucial role in standardizing ASL while they also contribute to regional dialect formation. As students from different backgrounds interacted, they developed distinct signing styles influenced by geography, education, and social networks. Dialectal differences appear in handshapes, movement, and sign choice, similar to regional accents in spoken languages. While standardization efforts exist, ASL continues to evolve naturally. These variations enrich ASL but also pose challenges for developing universal sign language recognition systems. Understanding dialect formation helps ensure inclusivity in sign language education and technology development.

ASL frequently interacts with English, which leads to linguistic borrowing and code-switching. Fingerspelling is a common form of borrowing, where English words are spelled out using the ASL alphabet, often for names or technical terms. Some signers mix ASL with English-based signing systems like Signed English, creating a form of code-switching. Loan signs also emerge when English words are adapted into ASL, modified to fit ASL's phonological system. These interactions highlight the dynamic nature of language contact and bilingualism in the Deaf community. Studying these influences helps improve sign language translation and documentation.

ASL is a dynamic and fully developed language with its own grammar, phonology, and sociolinguistic variations. While it shares similarities with spoken languages, its visual modality introduces unique linguistic phenomena. ASL continues to evolve through historical influences, dialectal diversity, and contact with English. Researching ASL provides valuable insights into language processing, variation, and accessibility technologies. Recognizing ASL's complexity is crucial for improving sign language

education, interpretation, and AI-driven recognition systems. As research advances, it contributes to linguistic equity and enhances communication opportunities for Deaf individuals worldwide.

2.2. Characteristics of KSL[2)]

Korean Sign Language (KSL) is a distinct visual-gestural language used by the Deaf community in South Korea. Though it shares some historical influences with Japanese Sign Language (JSL) due to Japan's occupation of Korea (1910-1945), KSL has developed independently over time. It possesses unique linguistic, structural, and sociolinguistic characteristics, making it a fully developed language rather than a derivative of spoken Korean. As a natural language, KSL plays a crucial role in communication, cultural identity, and education for Deaf individuals in Korea. Understanding its evolution and features helps promote linguistic diversity and accessibility for the Deaf community.

KSL has been in use since at least 1889 and was officially incorporated into Deaf education in Korea in 1908. During Japanese occupation, JSL was introduced into Korean schools, leading to some shared linguistic features between the two languages. However, KSL retained distinct elements and continued to evolve independently. Over the years, Deaf education and community interactions have played key roles in shaping the development of KSL. Today, it stands as a fully recognized language with its own grammatical structure and unique characteristics, distinct from both JSL and other sign languages worldwide.

KSL consists of key phonological components, including handshape, movement, location, palm orientation, and nonmanual signals such as facial expressions and body posture. Handshapes are essential in distinguishing meaning, while movement and palm orientation provide additional grammatical information. Some handshapes are exclusive to fingerspelling, which in KSL is based on Hangul's phonological principles. Unlike other East Asian sign languages, KSL's fingerspelling system mirrors the syllabic structure of written Korean. Nonmanual markers also play a vital role in conveying emotions and grammatical elements, making them an integral part of the language's overall structure.

KSL employs indexical classifiers (ICs) to indicate subjects or objects, marking its verb agreement system as distinct. Approximately one-third of KSL agreement verbs require classifiers, which are more flexible compared to JSL. Additionally, KSL utilizes specific

2) The summary in this section is primarily based on Fischer & Gong (2010).

handshapes "such as  and  " as morphemes to convey meaning, which represent concepts such as gender, kinship, and positive or negative connotations. This unique morphological system allows for more nuanced expression and highlights the complexity of KSL as a natural language.

KSL generally follows a Subject-Object-Verb (SOV) word order, similar to spoken Korean. However, topicalization allows the topic to be introduced at the beginning of a sentence, regardless of its grammatical role. Nonmanual markers contribute to KSL's syntax by marking negation, questions, and emphasis. Unlike ASL, where negation often extends across an entire clause, KSL allows negation through manual signs or more localized nonmanual markers. *Wh*-questions in KSL sometimes require distinct facial expressions that appear only at the end of the question, differing from Western sign languages where facial expressions often spread across the whole clause.

KSL exhibits regional variation due to historical factors. Differences in signing exist between Seoul and provincial areas, reflecting dialectal distinctions. Additionally, generational variation is evident, as older signers may use signs influenced by JSL, while younger signers favor newer KSL signs. This lexical variation reflects the natural evolution of the sign language and the influence of social and historical contexts. Understanding these variations is essential for standardizing sign language education and ensuring effective communication among different generations and regional communities within Korea's Deaf population.

KSL has been influenced by multiple languages throughout history. The Japanese occupation introduced JSL elements into KSL, some of which persist today. Additionally, KSL has borrowed words from spoken Korean, often adapting them through fingerspelling or initialized signs. In more recent years, some ASL signs have been introduced through international Deaf community interactions, though ASL's influence is less significant compared to JSL. These language contacts highlight KSL's adaptability and its interactions with other linguistic systems, shaping its lexicon and grammatical structure over time.

KSL has gained official recognition as a national language distinct from Signed Korean, which is a manually coded version of spoken Korean. Despite historical emphasis on oralism in Deaf education, KSL is now increasingly accepted in schools. It is also used in media, public services, and broadcasting to ensure accessibility for the Deaf community. The growing recognition of KSL highlights its importance as a linguistic and cultural identity marker for Deaf individuals in Korea. Expanding its presence in

education and media strengthens inclusivity and promotes linguistic rights for the Deaf community.

Korean Sign Language (KSL) is a fully developed sign language with unique linguistic features, including an SOV word order, classifier-based verb agreement, and a Hangul-influenced fingerspelling system. Shaped by historical events, Deaf education, and community use, KSL has evolved independently while maintaining some connections to JSL. It plays a vital role in South Korea's Deaf identity and cultural expression. As awareness of KSL grows, continued research and advocacy are essential to ensure its preservation, standardization, and wider acceptance in education, technology, and public life.

3. Sign Language Recognition

Sign Language Recognition (SLR) is the process of automatically identifying and interpreting sign language gestures using computational models. It plays a crucial role in bridging communication barriers between the deaf and hearing communities, by enabling sign-to-text or sign-to-speech translation. Unlike spoken languages which rely on linear sequences of phonemes, sign languages are highly dynamic: utilizing hand shapes, movements, palm orientations, facial expressions, and body posture to convey meaning. These elements introduce significant challenges in designing effective SLR systems.

Deep learning has revolutionized SLR by leveraging powerful models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to process visual and/or temporal information. Transformer-based architectures (Vaswani et al., 2017) further enhance performance by capturing long-range dependencies within signing sequences. However, recognizing continuous sign language remains a fundamental challenge due to the complexity of sign transitions, co-articulation effects, and signer variability. Unlike isolated word recognition, where each sign is manually segmented, continuous sign language recognition (CSLR) requires identifying individual signs within fluid and overlapping movements.

One of the major difficulties in CSLR is temporal segmentation, where signs transition seamlessly without explicit boundaries. The absence of clear segmentation cues leads to misinterpretations, particularly when two visually similar signs differ only by subtle motion differences. Additionally, non-manual markers, such as facial expressions and head

tilts, provide grammatical information in many sign languages, but they are difficult to model accurately due to variations in signer style and emotion. Spatial referencing, where signers use space to indicate subjects and objects, further complicates recognition as the same sign can have different meanings depending on spatial context.

To advance SLR systems, deep learning models must improve in capturing hierarchical features, integrating multi-modal inputs, and generalizing across different signers and dialects. Addressing these challenges will be key issues to developing robust, real-time sign language recognition systems that can facilitate seamless communication for the deaf community.

Cheng et al. (2020) introduces a Fully Convolutional Network (FCN) for continuous sign language recognition (CSLR). Unlike previous CNN-RNN hybrid models, FCN enables end-to-end training without pre-training. The model incorporates a novel gloss feature enhancement (GFE) module, improving sequence alignment by using rectified supervision and joint training. The focus is on learning individual glosses rather than complete sentence sequences, leading to better generalization and online recognition. The proposed FCN-based model outperforms previous CNN-RNN hybrid models, achieving state-of-the-art (SOTA) results on RWTH-PHOENIX-Weather-2014[3] and CSL datasets.[4] The study emphasizes efficiency and robustness in recognizing phrases, paragraphs, and signer pauses. However, the model relies on weakly annotated data and struggles with long-range dependencies. While the FCN-based model outperforms previous methods, achieving state-of-the-art (SOTA) results on the CSL dataset, it may face challenges when applied to KSL due to its classifier-based verb agreement system. Unlike ASL, KSL conveys grammatical relations primarily through movements and handshape classifiers, adding complexity to sequence alignment.

Slimane et al. (2021) proposes a self-attention-based network for continuous sign language recognition (CSLR). The model processes separate data streams for different sign language components, such as handshapes and facial expressions, to improve recognition

3) The RWTH-PHOENIX-Weather-2014 dataset is a benchmark dataset for German Sign Language (DGS) recognition. It contains video recordings of weather forecasts presented by sign language interpreters. The dataset includes continuous sign language sequences, gloss-level annotations, and spoken language translations.

4) The CSL (Chinese Sign Language) Dataset is a large-scale video dataset for continuous sign language recognition (CSLR). It includes various signers, sentence-level annotations, and multiple camera views. The dataset helps train AI models for gesture recognition and sign language translation, advancing research in Chinese Sign Language processing.

accuracy. It utilizes an attention mechanism to capture spatiotemporal dependencies and integrates handshapes with their surrounding context. The model achieves competitive performance on RWTH-PHOENIX-Weather 2014, demonstrating that attention mechanisms significantly enhance sign language recognition by incorporating contextual information. However, the approach requires large training datasets and mainly focuses on the dominant hand and facial expressions, which may not generalize well across different sign languages. For KSL, this method may encounter difficulties due to the complexity of non-manual markers. KSL relies more heavily on head movements and upper-body posture variations for grammatical structure than ASL, requiring models to incorporate a broader range of non-manual signals. Additionally, the importance of classifier handshapes in KSL means that simple attention mechanisms might struggle to distinguish subtle grammatical differences.

Renz et al. (2021) presents a sign language segmentation method using 3D convolutional neural networks (CNNs) and temporal refinement techniques. Unlike previous hand-tracking-based methods, this approach processes video frames directly, improving robustness in sign segmentation. The model is tested on British Sign Language (BSL) and German Sign Language (DGS) datasets, showing improved performance across different languages. It enhances segmentation accuracy, which is essential for reducing the cost of sign language dataset annotation and supporting downstream tasks like sign language translation. However, the approach relies heavily on dataset quality and struggles with signer variability. The presence of initialized signs using *Hangul* characters also adds complexity to automatic segmentation.

Rastgoo et al. (2021) introduces a multi-modal zero-shot learning (ZSL) framework for sign language recognition. The proposed method integrates visual and semantic information, overcoming traditional ZSL methods' reliance on visual features alone. The model projects both visual and textual representations into a shared multimodal embedding space, enabling recognition of unseen signs. Evaluations on benchmark datasets demonstrate improved ZSL performance compared to state-of-the-art methods. However, the model depends on high-quality semantic descriptions and requires further evaluation on diverse datasets to ensure generalizability. For KSL, adapting this framework would be challenging because of its unique phonological structure. The *Hangul*-based fingerspelling system introduces sub-syllabic components that do not directly correspond to English words, complicating semantic representation. Moreover, classifier verbs in KSL which depend on referential space usage, require additional adaptation in

the embedding space to maintain linguistic consistency.

Hu et al. (2021) presents SignBERT, a self-supervised pre-trained model for sign language recognition, focusing on hand movements. By masking parts of hand pose information and reconstructing them, the model learns to capture sign structure effectively. Unlike conventional methods that rely on visual features, SignBERT leverages hand poses, improving recognition accuracy across multiple datasets. Despite achieving state-of-the-art results, it may not fully capture non-manual markers such as facial expressions and upper-body posture. For KSL, this method might struggle because KSL relies on classifiers to encode spatial relationships and verb arguments, which hand pose alone may not fully capture. Additionally, fingerspelling in KSL follows a different structure than ASL's alphabet-based approach, making it difficult to integrate SignBERT's learned representations without extensive adaptation.

Bilge et al. (2022) introduces a zero-shot learning (ZSL) framework that enables the recognition of unseen sign classes by leveraging textual and attribute-based descriptions. Instead of training on visual data alone, the model incorporates sign language dictionary definitions to improve generalization. The framework is evaluated on ASL datasets and demonstrates that combining semantic descriptions with visual embeddings significantly enhances recognition accuracy. However, the approach is sensitive to the quality of textual descriptions and is primarily designed for recognizing isolated signs rather than continuous sign sequences. For KSL, this method would need adjustments because KSL signs often encode information through classifier predicates. Additionally, KSL's use of topic-comment structures means that isolated signs may not carry the same meaning without their contexts.

Madhiarasan et al. (2022) provides a comprehensive review of sign language recognition methods, covering various sign language types, data collection techniques, and existing datasets. It highlights the impact of factors such as signer variability, environmental conditions, and dataset biases on recognition accuracy. The review also discusses different deep learning architectures, identifying research gaps and potential future directions. However, the field is rapidly evolving, and certain aspects of the review may become outdated. For KSL, the review underscores the lack of large-scale datasets that include classifier-based verb agreement and *Hangul*-initialized signs. Many existing datasets primarily focus on Western sign languages, which have different morphological structures, making it difficult to directly apply the pre-existing deep learning models to KSL.

Akandeh (2022) introduces a sentence-level sign language recognition (SLR) framework using Connectionist Temporal Classification (CTC). The model processes continuous sign language sequences without requiring segmentation into individual words. Two models are tested: one using raw video frames and another incorporating hand shape and movement data extracted with Mediapipe. The second model achieves a lower word error rate (WER), demonstrating the benefits of incorporating hand feature data. However, the dataset is limited to weather reports, making generalization difficult. For KSL, this approach may be problematic due to its reliance on CTC, which assumes a left-to-right sequence. KSL's flexible syntax and classifier predicates mean that direct alignment with a sequential output model may not work as effectively without additional linguistic adaptations.

Lim et al. (2023) paper presents a lightweight deep learning model for ASL recognition on the humanoid robot Pepper. The system employs a Transformer encoder for sequential data processing while using large language models (LLMs) to generate co-speech gestures, making interactions more natural. The model achieves real-time ASL recognition on embedded systems, improving human-robot interaction. However, the system is limited by environmental conditions such as lighting and background noise. For KSL, adapting this system would require significant changes due to the structural differences between ASL and KSL. KSL's frequent use of two-handed classifier constructions and spatial referencing would require additional spatial modeling.

Moryossef et al. (2023) proposes a linguistically motivated sign language segmentation model using BIO tagging. It integrates optical flow features to capture prosodic cues and employs 3D hand normalization for improved representation of handshapes. The model performs well on German Sign Language (DGS) and generalizes across different sign languages. However, its reliance on pose estimation systems introduces potential errors. For KSL, segmentation would be complicated by its heavy reliance on classifier predicates. Additionally, topic-comment structures could make the phrase segmentation harder.

Rastgoo et al. (2024) introduces a Transformer-based model for detecting sign boundaries in continuous videos without relying on handcrafted features. The model achieves strong results on Persian and ASL datasets but struggles with variable sign durations and requires predefined frame numbers for accurate segmentation. For KSL, KSL's use of space for referential indexing would also require adjustments to the model's attention mechanisms to capture spatial dependencies.

4. Sign Language Production

Sign Language Production (SLP) refers to the process of automatically generating sign language gestures from spoken or written text using computational models. This technology is essential for improving accessibility and enabling real-time communication between deaf and hearing individuals. Unlike spoken languages, sign languages are highly visual and spatial, requiring the precise generation of hand movements, facial expressions, body posture, and spatial referencing to convey meaning accurately. Deep learning has significantly advanced SLP by leveraging neural networks to synthesize sign gestures from textual input, often using motion capture data or avatar-based animation techniques.

One of the most promising approaches in SLP is the use of virtual avatars to represent sign language. However, avatar-based SLP faces several critical challenges that make the generation of natural and expressive sign language extremely difficult.

First, the complexity of sign language articulation makes it challenging to generate smooth and realistic hand movements. Unlike spoken languages, which follow a linear structure, sign languages rely on simultaneous articulations; where hand shape, motion, and facial expressions work together to convey meaning. Deep learning models often struggle to integrate these elements in a fluid and natural manner.

Second, avatars lack the subtle nuances of human signers, particularly in conveying emotions and prosodic features. Facial expressions and eye gaze play a crucial role in sign language grammar, such as marking questions, negations, or topic emphasis. Current deep learning models struggle to generate facial expressions that appear natural and contextually appropriate.

Third, ensuring accurate spatial referencing is difficult. Many sign languages use the signing space to represent subjects and objects dynamically. However, avatars often fail to maintain consistent spatial alignment, leading to unnatural or ambiguous signing.

To improve avatar-based SLP, future research must focus on refining motion synthesis techniques, enhancing non-manual gesture generation, and developing more realistic 3D avatars that can capture the full expressiveness of human signers. Solving these challenges is crucial for creating high-quality, accessible sign language production systems that truly benefit the deaf community.

Rastgoo et al. (2021) provides a comprehensive review of advances in Sign Language Production (SLP) using deep learning techniques. It categorizes studies based on input

modalities, datasets, applications, and model architectures, analyzing strengths and limitations. The review highlights challenges such as translating between different linguistic structures, generating photorealistic signers, and incorporating non-manual features like facial expressions. The paper discusses major approaches, including avatar-based models, neural machine translation, motion graph techniques, and conditional video generation. It emphasizes the need for large, high-quality datasets and suggests future research directions such as integrating multi-channel information and Graph Neural Networks. Despite advancements, generating realistic, high-resolution sign language videos remains challenging. For KSL, SLP faces additional difficulties due to its classifier-based verb system and *Hangul*-influenced fingerspelling. Unlike ASL, KSL often encodes grammatical relations spatially, making direct translation from spoken Korean more complex.

Jiang (2022) introduces SDW-ASL, a system designed to generate large-scale datasets for American Sign Language (ASL) production. The system uses deep learning and crowdsourced content selection from online news channels to extract human articulations in a condensed body pose format. The initial dataset consists of 30,000 sentences, 416,000 words, and a vocabulary of 18,000 words, totaling 104 hours of video. It is the largest continuous ASL dataset to date, aiming to facilitate deep learning research in ASL production. The approach ensures scalability while reducing manual annotation costs, but it relies on automated extraction techniques, which may introduce errors. For KSL, applying this system would be challenging due to differences in linguistic structure and representation. KSL's verb agreement and classifier system require spatial tracking beyond simple body pose extraction.

5. Sign Language Translation

Sign Language Translation (SLT) refers to the process of automatically converting sign language into spoken or written language and vice versa using deep learning models. This technology is crucial for bridging communication gaps between the deaf and hearing communities, promoting accessibility in education, healthcare, and public services. Unlike spoken language translation, which deals with linear sequences of words, SLT must process complex visual-spatial expressions, facial cues, and grammatical structures unique to sign languages. Recent advancements in deep learning, particularly transformer-based

architectures, have significantly improved sign-to-text and text-to-sign translation systems by leveraging large-scale datasets and multi-modal learning techniques.

The concept of a barrier-free society emphasizes equal access to communication, information, and services for all individuals, regardless of disabilities. SLT plays a fundamental role in achieving this vision by enabling seamless interaction between signers and non-signers. Without effective translation systems, the deaf community faces significant challenges in everyday communication, often relying on human interpreters, which may not always be available. Automated SLT offers a scalable and efficient solution to remove these barriers, ensuring inclusivity in workplaces, public institutions, and digital platforms.

However, automatic SLT remains a highly challenging task due to the structural differences between sign languages and spoken languages. Sign languages are not direct translations of spoken languages; and they follow distinct syntactic rules, such as topic-comment structures, classifier-based descriptions, and spatial referencing. Additionally, regional variations and personal signing styles make it difficult to develop models that generalize across different signers and dialects.

To advance SLT and move closer to a truly barrier-free communication system, deep learning models must improve in handling sign language grammar, integrating contextual meaning, and capturing real-time signing variations. Overcoming these challenges will allow SLT to become a key technology in fostering accessibility and inclusivity worldwide.

Rastgoo et al. (2021) reviews advancements in sign language translation (SLT), analyzing deep learning-based approaches. It categorizes existing methods based on input modalities, dataset types, and model architectures. It also highlights challenges such as syntactic differences between sign and spoken languages, the need for large-scale annotated datasets, and the difficulty of generating grammatically accurate translations. The paper discusses end-to-end neural machine translation models, transformer-based approaches, and techniques incorporating gloss-level annotations to improve translation accuracy. The authors suggest integrating multi-modal features such as hand movement, facial expressions, and spatial referencing to enhance translation models. Despite progress, generating coherent and contextually appropriate translations remains difficult due to the inherent visual-spatial nature of sign languages. For KSL, these approaches may face additional challenges due to its classifier-based verb agreement system and SOV word order, which differ significantly from English or ASL. Unlike ASL, KSL relies heavily on spatial relationships and role shifting, which current SLT models might struggle to encode

properly without specialized adaptations.

Jiang (2022) introduces SDW-ASL, a system designed to generate large-scale datasets for American Sign Language translation. The system uses crowdsourced content selection and deep learning techniques to extract human articulations in a condensed body pose format. The dataset is aligned with English text using closed captions from online news sources, ensuring structured annotation. This dataset aims to facilitate natural language processing (NLP) and machine translation research for ASL by providing sufficient data for deep learning models. However, it relies on news-based content, which may not reflect the full range of ASL expressions. For KSL, constructing a comparable dataset poses challenges due to dialectal variations and classifier-based signs, which often lack clear one-to-one mappings with written text. Furthermore, the dataset must account for KSL's *Hangul*-influenced fingerspelling system, which includes sub-syllabic components that differ significantly from English phonetic structures.

Lim et al. (2023) explores transformer-based architectures for sign language translation, proposing an attention-enhanced SLT model that integrates temporal and spatial context for improved performance. Unlike traditional gloss-based methods, the model learns directly from video data, capturing both manual and non-manual features to enhance translation accuracy. It employs a multi-stream transformer network, which processes facial expressions, hand gestures, and body movements separately before fusing them for final translation output. The model is evaluated on RWTH-PHOENIX-Weather and How2Sign datasets, achieving state-of-the-art performance in sentence-level sign language translation. Despite these improvements, the system struggles with long-range dependencies and ambiguous sign sequences, requiring further optimization. For KSL, this approach would need adjustments due to classifier-based verb agreement and spatial referencing, which require non-linear dependencies beyond simple temporal attention.

Moryossef et al. (2023) investigates zero-shot sign language translation, proposing a pre-trained model that leverages multilingual NLP techniques to translate sign language into spoken text without requiring language-specific training. The model aligns sign language features with a shared semantic space, allowing transfer learning from spoken language datasets. The study shows promising results in cross-lingual sign language processing, demonstrating that models trained on ASL can generalize to other sign languages with limited fine-tuning. However, challenges remain in handling linguistic structures unique to sign languages, such as spatial agreement, role shifting, and non-manual markers. The authors suggest incorporating self-supervised learning techniques

to improve generalization across different sign languages. For KSL, a zero-shot approach may struggle with its unique sentence structure and classifier system, which have no direct equivalents in spoken languages. Additionally, the influence of Hangul in KSL fingerspelling makes phoneme-level alignment difficult. These challenges require alternative tokenization strategies to align with multilingual NLP models, which can be addressed through subword vector techniques used in input encoding, such as byte-pair encoding (BPE) or WordPiece, enabling the model to handle morphologically rich and structurally diverse inputs more effectively.

6. Discussion

Sign Language Recognition (SLR) research has advanced significantly, but applying existing models to KSL presents unique challenges. Unlike ASL, KSL relies heavily on classifier predicates and spatial referencing, which are difficult to model using traditional sequence-based approaches. Many deep learning models, such as CNN-RNN hybrids or Transformer-based architectures, assume a linear structure. Additionally, KSL's topic-comment sentence structure and flexible word order complicate direct sequence-to-sequence translation models, requiring alternative approaches that can dynamically adjust to spatial referents. Another key issue is non-manual markers, such as head tilts and eyebrow movements, which serve grammatical functions in KSL but are often overlooked in SLR models trained on Western sign languages. Addressing these challenges will require multi-modal learning strategies and larger annotated datasets tailored specifically for KSL.

Sign Language Production (SLP) models face significant challenges when applied to KSL due to its unique grammatical structures and linguistic features. Unlike ASL, KSL relies heavily on classifier predicates and spatial referencing, which require dynamic space utilization rather than simple linear translations. Existing avatar-based SLP systems often struggle to maintain consistent spatial alignment, making accurate representation of KSL grammar difficult. Additionally, KSL's *Hangul*-influenced fingerspelling system differs from ASL's alphabet-based approach. Non-manual markers, such as head movements and facial expressions, also play a crucial grammatical role in KSL, but current deep learning models fail to generate natural and expressive non-manual gestures. Moreover, the two-handed signing system in KSL introduces further complexity in generating fluid

motion transitions. Future research must focus on multi-modal gesture synthesis, improved spatial alignment, and enhanced non-manual expression generation to develop more accurate and natural KSL production models.

Sign Language Translation (SLT) has advanced significantly through deep learning, yet its application to KSL presents unique challenges. Unlike ASL, KSL employs classifier-based verb agreement and spatial referencing, requiring models to process complex spatial relationships rather than simple sequential text alignment. Many existing SLT models rely on gloss-based annotation, which may not effectively represent KSL's topic-comment structure and flexible word order (SOV). Furthermore, *Hangul*-influenced fingerspelling introduces phonemic structures absent in ASL, making direct adaptation of pre-trained models difficult. Additionally, non-manual markers, such as eyebrow movement and head tilting, play grammatical roles in KSL but are often underrepresented in datasets. To improve SLT for KSL, multi-modal deep learning approaches must be explored; integrating spatial tracking, classifier predicates, and non-manual elements while addressing dialectal variations. Overcoming these challenges will be essential for developing accurate, real-time KSL translation systems that ensure accessibility in communication.

7. Conclusion

Sign Language Recognition, Production, and Translation (SLR, SLP, and SLT) have made significant progress with deep learning, yet adapting these technologies to KSL remains a complex challenge. Compared with ASL, KSL more relies on classifier predicates, spatial referencing, and non-manual markers, which make conventional sequence-based models insufficient.

Current SLR systems struggle with KSL's flexible word order and classifier-based verb agreement, requiring models that can dynamically process spatial meaning rather than relying on linear representations. Similarly, SLP models may encounter challenges in generating fluid, natural gestures, particularly in maintaining spatial accuracy and expressing non-manual grammatical markers, such as head movements and facial expressions. SLT systems must also evolve beyond gloss-based annotations to accurately reflect KSL's topic-comment structure and *Hangul*-influenced fingerspelling system, which differs from ASL's alphabetical system.

Addressing these challenges requires essentially larger annotated KSL datasets. I hope that future research should focus on integrating hand shape, motion, and facial expressions, refining gesture synthesis techniques, and developing real-time translation models to improve accessibility for KSL users. Advancing these technologies is essential for creating more inclusive communication tools, ensuring that KSL speakers benefit from the same level of accessibility as users of other sign languages.

For KSL, we recommend developing multi-modal transformer-based architectures that explicitly integrate spatial attention mechanisms to capture the three-dimensional signing space. Such models should combine 3D pose estimation for hand and body keypoints with high-resolution facial expression tracking to handle non-manual markers. Incorporating graph neural networks can further model spatial relationships between hands, body, and facial components, while sequence-to-sequence modules can adapt to flexible word order and classifier-based verb agreement. Additionally, pretraining on large-scale gesture datasets, followed by fine-tuning on KSL-specific corpora, will improve both recognition and production accuracy, enabling more natural and contextually appropriate translation outputs.

References

Akandeh, A. (2022). Sentence-level sign language recognition framework. arXiv Preprint arXiv:2211.14447.

Bilge, Y., Cinbis, R., & Ikizler-Cinbis, N. (2022). Towards zero-shot sign language recognition. arXiv Preprint arXiv:2201.05914.

Camgoz, N., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: joint end-to-end sign language recognition and translation. arXiv Preprint arXiv:2003.13830.

Cheng, K., Yang, Z., Chen, Q., & Tai, Y. (2020). Fully convolutional networks for continuous sign language recognition. arXiv Preprint arXiv:2007.12402.

Fang, B., Co, J., & Zhang, M. (2018). DeepASL: Enabling ubiquitous and non-Intrusive word and sentence-level sign language translation. arXiv Preprint arXiv: 1802.07584.

Fischer, S., & Gong, Q. (2010). Variation in East Asian sign language structures. In D. Brentari (Ed.), *Sign languages* (pp. 499-518). Cambridge: Cambridge University Press.

Hu, H., Zhao, W., Zhou, W., Wang, Y., & Li. H. (2021). SignBERT: Pre-training of hand-model-aware representation for sign language recognition. arXiv Preprint arXiv:2110.05382.

Jiang, Y. (2022). SDW-ASL: A dynamic system to generate large scale dataset for continuous American sign language. arXiv Preprint arXiv:2210.06791.

Ko, S., Kim, C., Jung, H., & Cho, C. (2019). Neural sign language translation based on human keypoint estimation. arXiv Preprint arXiv:1811.11436.

Lim, J., Sa, I., MacDonald, B., & Ahn, H. (2023). A sign language recognition system with pepper, lightweight-Transformer, and LLM. arXiv Preprint arXiv:2309.16898.

Lucas, C., & Bayley, R. (2010). Variation in American sign language. In D. Brentari (Ed.), *Sign languages* (pp. 451-475). Cambridge: Cambridge University Press.

Madhiarasan, M., & Roy, P. (2022). A comprehensive review of sign language recognition: Different types, modalities, and datasets. arXiv Preprint arXiv: 2204.03328.

Moryossef, A., Jiang, Z., Muller, M., Ebling, S., & Goldberg. Y. (2023). Linguistically Motivated Sign Language Segmentation. arXiv Preprint arXiv:2310.13960.

Rastgoo, R, Kiani, K., Escalera, S., & Sabokrou, M. (2021). Sign language production: A review. arXiv Preprint arXiv:2103.15910.

Rastgoo, R., Kiani, K., & Escalera, S. (2024). A transformer model for boundary detection in continuous sign language. arXiv Preprint arXiv:2402.14720.

Rastgoo, R., Kiani, K., Escalera, S., & Sabokrou, M. (2021). Multi-modal zero-shot sign language recognition. arXiv Preprint arXiv:2109.00796.

Renz, K., Stache, N., Albanie, S., & Varol, G. (2021). Sign language segmentation with temporal convolutional networks. arXiv Preprint arXiv:2011.12986.

Slimane, F., & Bouguessa, M. (2021). Context matters: Self-attention for sign language recognition. arXiv Preprint arXiv:2101.04632.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin I. (2017). Attention is all you need. arXiv Preprint arXiv: 1706.03762.

Yin, A., Zhao, Z., Liu, J., Jin, W., Zhang, M., Zeng, X., & He, X. (2021). SimulSLT: End-to-End simultaneous sign language translation. arXiv Preprint arXiv: 2112.04228.

Yong-hun Lee
Research Professor
Department of Linguistics
Chungnam National University
99 Daehak-ro, Yuseong-gu, Daejeon 34134, Republic of Korea
Phone: +82-42-821-5318
E-mail: yleeuiuc@cnu.ac.kr

Received on April 13, 2025
Revised version received on June 23, 2025
Accepted on June 30, 2025