Revisiting Sign Language from an NLP Perspective

Anonymous ACL submission

Abstract

This paper argues that current Natural Language Processing (NLP) frameworks are fundamentally misaligned with sign language processing (SLP) due to their reliance on linear, single-channel linguistic models commonly used in spoken language processing. We analyze the fundamental differences between spoken and signed languages across four critical dimensions: (1) multi-modal vs. multi-channel representation, (2) low-resource vs. high-resource data availability and annotation efficiency, (3) disambiguation vs. channel conversion, and (4) linearity vs. spatiality representation. Existing research primarily focuses on surface-level forms, neglecting deep semantic structures that rely on coordinated multi-channel features inherent to sign languages. We identify three underexplored challenges that highlight these gaps: the spatial modeling challenges of text-to-scene conversion, the dual representation problem in spatial metaphors, and the complexity of classifier predicate decomposition. These challenges demonstrate that SLP cannot be reduced to video-to-text or text-to-video translation, and instead requires a fundamental rethinking of NLP's core assumptions to integrate the spatial-semantic structures of sign languages.

1 Introduction

Sign language is an important means of communication for many deaf individuals and other users. It is not merely a simple combination of signs but a complete and complex natural language with multi-channel encoding properties (Valli, 2000). However, in the field of NLP, studies on sign language modeling receive limited attention (Yin et al., 2021). Recent research on Sign Language Translation (SLT) and Generation (SLG) has increased. However, most studies focus on surface-level representations, such as gesture extraction and corresponding annotation. The focus on the linguistic organization and deep semantic structures of sign language—particularly its spatial semantics, and multi-channel interactions—is still insufficient. This limitation has led to the status of sign language in NLP research being far from equal to that of other spoken languages, and it also fails to fully meet the growing communication needs and technological expectations of the Deaf community.

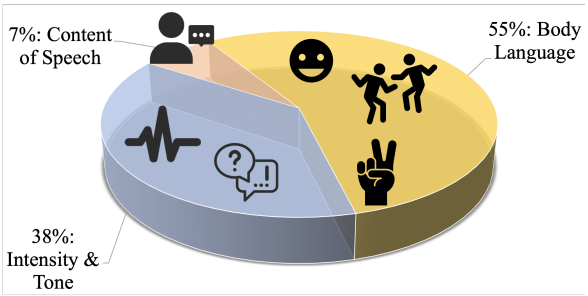

Figure 1: The Impact of Non-verbal Cues in Human Communication: Speech Content, Intensity and Tone, and Body Language.

Despite the availability of sign language resources, effective methods for exploration and utilization remain lacking. Yao et al. (2019) described these resources as underutilized and difficult to access. Currently, sign language processing (SLP)[1] is still constrained by corpus collection and manual annotation, limiting progress in knowledge organization, classification, and deeper linguistic analysis.

Sign language, like spoken and written language, is a product formed from human perception, experience, and cognitive processing of the objective world. It allows linguistic features to be analyzed independently of physical expression, contributing to research on spatial cognition and its impact on semantic structures, including the essence of language, its structure, language acquisition by children, and cognitive mechanisms in the human brain

[1] SLP refers to Sign Language Processing in this paper, except where explicitly noted as Sign Language Production. Context clarifies usage.

(Aronoff et al., 2020). As shown in Figure 1, cognitive psychology research suggests that in natural language communication, speech content accounts for only 7% of the impact, while intensity and tone contribute 38%, and body language 55%, respectively (Mehrabian, 1973). Despite these findings, computational linguistics has primarily focused on speech content, with less attention to non-verbal communication such as facial expressions and body movements (Li, 2024). Li further emphasized that non-verbal behaviors are primary channels for information transmission in human interaction.

A common scenario might be an old friend claiming to be overwhelmed with work, yet his relaxed expression implies otherwise. This discrepancy between speech content and body language cues reflects sign language's multi-channel structure, where manual (handshape, movement, location, orientation) and non-manual (facial expressions, torso movement) features convey layered meaning simultaneously. In contrast, spoken language usually spans multiple modalities—auditory plus possible visual signals—whereas sign language remains within a single visual modality but orchestrates multiple channels at once. Consequently, traditional NLP, optimized for linear text or speech, struggles to handle sign language's synchronized spatial data.

Research on sign language enhances the understanding of non-verbal communication and contributes to the development of AI systems capable of interpreting such communication. This paper systematically contrasts spoken and sign language processing across four critical dimensions, identifies three unresolved theoretical challenges, and proposes pathways to bridge these gaps. Our analysis aims to catalyze NLP research that respects sign language's linguistic integrity while advancing inclusive technologies.

2 Background and Related Work

At present, merely capturing and collecting sign language videos for recognition and understanding cannot be considered true informatization. To achieve informatization, computers must be capable of performing a range of operations, including digitizing, analyzing, storing, transmitting, and presenting sign language. Although sign language has established writing systems, such as the SignWriting system in the United States (da Rocha Costa et al., 2003), which enables computer encoding

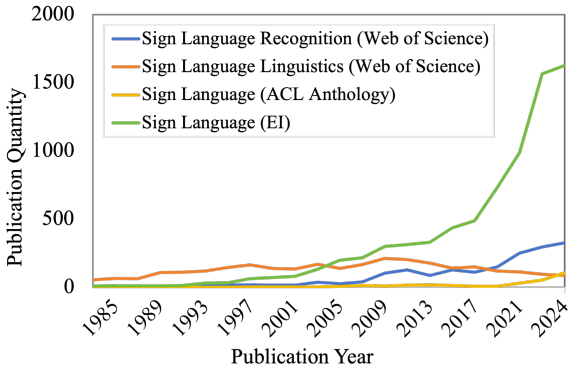

Figure 2: Trend of Sign Language-Related Publications (1983–2024)

for storage, transmission, and display, these systems have a limited audience, and information processing based on them has not yet been widely adopted. This disconnection between linguistic notation and computational processing stems from sign language's multi-channel nature.

Sign language information processing research dates back to 1983, and our statistical analysis covers works published until December 31, 2024. The ACL (Association for Computational Linguistics) Anthology contains approximately 540 sign language-related papers, with publications dating back to 2003. In the EI (Engineering Index) database, there are about 13,176 sign language-related papers, while a search in Web of Science yields 3,122 papers with "Sign Language Recognition" as a keyword and 5,480 papers with "Sign Language Linguistics" as a keyword. Figure 2 illustrates the annual publication trends for these papers.

In comparison to the hundreds or thousands of papers published annually in these two fields—Sign Language Recognition (SLR) in Computer Vision (CV) and sign language linguistics in Linguistics—the literature on sign language-related NLP is relatively limited. As Yin et al. (2021) emphasized, it is now time to apply insights from computational linguistics to the modeling of sign language.

In the field of CV, there is an abundance of literature on SLR, which will not be reiterated here. Generally speaking, such studies have made certain progress in recognition accuracy, partially leveraging language models to improve semantic understanding. However, most SLR models still lack an in-depth analysis of sign language linguistics and often treat it as a "sequence of videos + glosses"

model, achieving only surface-level matching (Abdullah et al., 2024). As a result, although some preliminary SLR or SLT systems have been deployed, a commercialized solution widely accepted by the deaf community has yet to emerge. The primary reasons include the relatively small size of the databases, limited vocabulary, insufficient language coverage, and suboptimal semantic accuracy and fluency in real-world scenarios. More importantly, the multi-channel nature inherent in sign language has yet to be fully utilized, making it challenging to capture crucial information such as facial expressions and body movements beyond the visual modality. This highlights that relying solely on computer vision for "video-to-text" or "text-to-video" conversion is insufficient to meet the diverse needs of sign language in practical applications.

There is also a wealth of traditional research literature on sign language linguistics, most of which focuses on the linguistic structure and semantic layers of sign language itself. The use of computational tools for sign language research has become the mainstream approach. Several corpus projects such as the Public DGS Corpus (Hanke et al., 2020), ECHO Corpus (Kopf et al., 2022), and Auslan Corpus (Cassidy et al., 2018) often utilize multi-level transcription tools like ELAN (2024), iLEX (Hanke and Thomas, 2002), and SignStream (Neidle et al., 2018)for fine-grained annotation of sign language expressions. The NCSLGR (National Center for Sign Language and Gesture Resources Corpus) at Boston University (Neidle and Vogler, 2012), for example, has conducted multilayer analyses of sign language videos, making it an informative linguistic resource. However, these annotated corpora are primarily aimed at pure linguistic research, emphasizing annotation accuracy and systematicity. On the one hand, they involve high levels of manual participation, making them costly and time-consuming. On the other hand, these fine-grained annotations do not account for the needs of information processing, making it difficult to scale to large data scenarios efficiently. As a result, although these corpora have significant academic value, there remains a gap between them and practical applications such as SLR, SLT, and SLG. Consequently, research outcomes are challenging to scale and apply in real-world contexts.

From an NLP perspective, studying sign language requires addressing both its dual role as a visual and linguistic system and the challenges of deep semantic analysis. This requires transcending pure computer vision approaches to engage with linguistic structures across phonological, morphological, syntactic, and semantic layers. Particularly crucial is semantic understanding, which enables accurate interpretation of signers' communicative intent. Our research re-examines sign language's linguistic features from this NLP oriented perspective, shifting focus from traditional video-level gloss extraction to multi-dimensional semantic understanding and generation.

3 A Comparison of Sign Language and Spoken Language from the Perspective of NLP

The differences in NLP between spoken and sign languages are not just due to the lack of writing system-related foundations for sign language, nor can they be simplified to a one-to-one translation. The core issue is that NLP theories are based on single-channel processing, while sign language is multi-channel, making the application of spoken language NLP theories to sign language complex. The modality of spoken language is typically carried by speech, which is a set of values that change over time (Huenerfauth, 2005).

The writing system of spoken language functions similarly, requiring only the recording of written symbols that correspond to speech. Both written symbols and speech data are time-series streams confined to a single channel. This temporal structure forms the foundation of spoken language NLP systems. In contrast, sign language is characterized by its spatiality, giving it multi-channel nature. This complexity makes encoding sign language into a linear single-channel format inherently difficult, resulting in the loss of crucial linguistic details during processing. Sign language linguists (Sandler, 2012; Brentari, 2019) have identified that manual features, including handshape, location, orientation, and movement, as well as non-manual features, such as facial expressions and torso movements, all convey essential linguistic meaning. These channels are interdependent and inseparable. This distinct linguistic mode, unlike the linear nature of spoken language, complicates sign language processing, leading to significant differences in computational processing for spoken and sign languages.

3.1 Spoken Language Multi-modal vs. Sign Language Multi-channel

With the advancement of large models such as ChatGPT and Sora in the market, there has been a shift from single-modal to multi-modal capabilities, with multi-modal research gradually becoming mainstream. Multi-modal research involves integrating and analyzing information from multiple sensory modalities(e.g., text, speech, and vision) to understand and generate cross-modal content, thereby better simulating human cognition—how humans understand the world through multiple senses (Fei et al., 2024). However, even in a multi-modal environment, traditional NLP for spoken language typically focuses only on single-channel speech or single-channel text processing, without the need to account for such a wide range of additional channel information.

In contrast, sign language is not only visual but also inherently multi-channel, meaning that it conveys information through multiple coordinated expressive channels within the same modality. Multi-channel processing in sign language uses various channels within one modality to convey information. For example, eyebrow changes, head movements, body postures, and lip movements can all contribute to the communication of complex linguistic details. These multi-channel interactions operate simultaneously, fundamentally differing from the multi-modal processing of spoken language.

Thus, multi-channel processing emphasizes coordinating different channels within one modality, whereas multi-modal processing combines information across various modalities for deeper insights and more accurate understanding. The complexity of sign language, with its multi-channel nature, requires researchers to account for these channels in processing, adding to the technological development challenges. Deep research into sign language's multi-channel nature could contribute to improving and enhancing NLP systems.

3.2 Spoken Language High-Resource vs. Sign Language Low-Resource

The progress of large models today is largely driven by the availability of vast digital data sources and the development of large-scale automatic annotation tools. These resources play a crucial role in transforming unstructured data into more structured forms, which in turn supports the creation of rich training corpora for machine learning. This has contributed to significant advancements in pre-trained models and Large Language Models (LLMs). In contrast, although a substantial number of sign language videos are available online, many remain unstructured and lack detailed annotations, making them difficult to leverage effectively for ML due to the high cost of manual annotation. The Real-Time Factor (RTF) for annotating spoken language is 1, meaning one hour of annotation equals one hour of work. However, the RTF for sign language can reach as high as 100, meaning that one hour of sign language corpora requires 100 hours for annotation (Dreuw et al., 2008). This stark difference stems from the single-channel nature of spoken language compared to the multi-channel nature of sign language, significantly increasing the complexity and cost of developing automatic annotation technologies.

The multi-channel nature of sign language demand the integration of knowledge from multiple disciplines, including CV, ML, NLP, and sign language linguistics. This poses significant challenges for researchers, requiring a high level of interdisciplinary expertise, especially a deep understanding of sign language linguistics (Bragg et al., 2019). These challenges present major obstacles to the development of automatic annotation technologies, leaving sign language in a "low-resource" state in the current landscape.

Recent advancements have led to sign language LLM, but limitations exist due to insufficiently large and detailed annotated data. These models depend on videos or surface data like keypoints, which hinders a full understanding of sign language's linguistic features. Consequently, they lack the ability to fully understand and reason at syntactic, semantic, or pragmatic levels, resulting in sign language LLM not yet matching spoken language LLM in language understanding and generation capabilities.

3.3 Disambiguation in Spoken Language vs. Channel Conversion in Sign Language

The fundamental task of spoken NLP is disambiguation across morphology, syntax, and semantics. In sign language NLP, it's also a task but not the core. High uncertainty in one channel, such as manual features, can be reduced by others, like facial expressions and body movement. Deaf individuals need less phonetic information to recognize a single gesture, faster than spoken words. This information is constrained by the phonological structure

of sign language and the early and simultaneous availability. Both are aiding rapid recognition.

Gestures, as visual signals, inherently provide early synchronous phonological information. Typically, after about 150 ms of a gesture, its location and palm orientation can be recognized, and after about 20 ms, the handshape can be identified (Emmorey and Corina, 1990). This early availability narrows down possible gesture candidates, aiding faster recognition. Additionally, sign language's phonological and morphological structures differ from spoken words. The garden path phenomenon, common in spoken language ambiguity, is rarely observed in sign language, possibly due to simultaneous visual cues (Huang and Ferreira, 2021). In spoken language, phonetic overlap is common (e.g., over 30 words share the sounds [kan], [mæn], and [skr]), whereas in sign language, multiple gestures rarely share the same initial handshape and target location. This phonological structure also limits the size of the initial list of sign candidates (Emmorey, 2001). Such visual cues help Deaf individuals predict a gesture's morphological structure (Fine et al., 2005; Emmorey et al., 2009).

Disambiguation generally aims to reduce uncertainty in language understanding and is often linked to information entropy. However, entropy (unpredictability) and redundancy (contextual cues resolving ambiguity) are distinct. Chinese, as one of the spoken languages, is widely recognized as one of the most concise languages, with a greater information entropy (Montemurro and Zanette, 2001). Therefore, disambiguation in Chinese is more costly and less efficient than in other spoken languages, as it relies more heavily on pragmatic knowledge, such as context and world knowledge. There is no relevant literature on the information entropy of sign language, but according to our self-constructed sign language corpus, the maximum length of a single gesture is equivalent to 11 Chinese words, or roughly 18 Chinese characters (see Figure 3). From this, we infer that sign language's information entropy is higher than spoken languages, with variations across different channels. Non-manual channels like facial expressions and body movements act as pragmatic knowledge, providing more context-based information than spoken language. This suggests sign language may have greater redundancy, reducing entropy and uncertainty. Calculating its entropy requires a large corpus and further experiments.

Compared to spoken language NLP, the core task in sign language NLP is the conversion between single-channel and multi-channel features. Current theories in spoken language NLP mainly focus on computing the average codeword length for a single channel, with little attention given to multi-channel systems. Shannon's first theorem (1948) states that the average length of a codeword can only be greater than or equal to the entropy of the information source. Spoken NLP optimizes codewords under one-channel capacity constraints; sign language, by contrast, must handle high-entropy, multi-channel inputs. This complexity hinders progress in sign language processing, making it less efficient than spoken language.

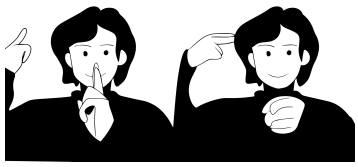

Figure 3: This is a sign in Chinese Sign Language. One hand forms the "command" sign (pointing to oneself), while the other hand makes the "guiding" sign (gesturing outward). This sign is typically used when the referents have already been established in the conversation. The phrase it conveys is "在A的控制下, 他成为某国的B傀儡" (Under the control of A, he served as the puppet emperor of a certain country).

It is urgent to address the input-output challenges in sign language and advance the theory of multi-channel coding. The focus of NLP research should gradually shift toward multi-channel coding, aligning the theories of spoken language NLP and multi-channel coding.

3.4 The Linearity of Spoken Language vs. the Spatiality of Sign Language

A fundamental limitation of traditional NLP lies in the inherent linearity of spoken language, which cannot simulate the spatial layout of entities and objects in the three-dimensional (3D) scenes of sign language. Huenerfauth (2004) argues that translating spoken language into sign language requires simulating a signer's mental 3D space, mapping entities referenced in the spoken language to this mental space, and subsequently mapping them to the physical space through gestures, thereby conveying the meaning of the source text. For instance, translating "The car is next to the house" in sign language requires selecting a classifier handshape for the car (e.g., four wheels), placing it relative to the

house within the signer's two-handed space, and integrating non-manual cues (e.g., furrowed brows, torso shifts) to depict motion or context. This highlights that moving from single-channel (spoken) to multi-channel (signed) language involves spatial representation and often demands contextual knowledge(common sense, world knowledge).

Currently, text-to-scene processing in spoken language NLP primarily focuses on the spatial deployment of entities and has yet to fully address multi-channel transformation processes. In contrast, sign language NLP incorporates more refined spatial concepts, necessitating the supplementation of information absent in spoken language to accomplish the fundamental task of spatial conceptual transformation.

The signing space in front of a signer's body represents different meanings. Sutton-Spence and Woll (1999) proposed dividing the signing space into topographical space and syntactic space. Topographical space refers to mapping the positions of objects in actual space to corresponding positions in the signing space. It is typically used to describe the positions and motion directions of persons or objects. For example, when describing a driving scene, the sign for "car" can be made while moving through the space in front of the signer to depict the motion. To express "winding roads," the sign can change direction back and forth; to indicate "bumpy roads," the sign can move up and down. In this way, sign language naturally completes the description of real-world spatial relationships. A study comparing the cognitive processing differences between these two types of spaces found that Deaf individuals responded faster to judgment tasks after viewing sentences involving topographical space than to those involving syntactic space (Hickok et al., 1996). This finding suggests that Deaf individuals process these two types of spaces differently. Consequently, spatial computing is an essential topic in sign language NLP that cannot be overlooked.

In the morphological stage, spatial modeling uses non-actual space, especially in pronouns, verbs, and comparative gestures, adjusting hand direction based on subject-object positions.

In pronoun usage, signers refer to entities using a position close to the torso, where distance indicates relationships. This allows space to function as pronouns, differing from spoken language's single-channel limitation. Sign language uses multiple channels, enabling infinite subdivisions of referents. Consequently, pronouns can refer to specific entities, making spatial references vivid but potentially confusing with multiple referents.

In syntax, sentences use real space through spatial verbs and classifier predicates, integrating linguistic and spatial features. Classifier predicates do not rely on spoken language's spatial prepositions, instead constructing multi-channel representations through spatial scene depictions. This transformation exemplifies sign language's flexibility in 3D spatial descriptions.

4 Theoretical Issues to Be Resolved

While theoretical models such as phonological classification and verb categorization have contributed to sign language NLP, core challenges remain. These include spatial modeling, the representation of spatial metaphors, and understanding the cognitive mechanisms of classifier predicates. A comprehensive theoretical framework has yet to be fully established. Although some sign language NLP research has introduced new technologies—such as digital humans, data gloves, and wearable devices—to develop 3D signer models and sign language corpora, much of the work still relies on adapting traditional NLP methods, which may not fully address sign language's unique spatial and multi-channel nature. Several theoretical issues, therefore remain to be resolved.

4.1 Text-to-Scene Conversion

Similar to text-to-scene conversion in spoken language, SLG emphasizes the modeling of spatial relationships between objects. It is restricted to spatial layouts and does not require the sequencing of images or the modeling and placement of related conceptual entities. Text-to-scene conversion is a novel research topic in NLP with limited studies available. When applied to sign language NLP, it encounters challenges related to linguistic ambiguity and the unclear expression of spatial concepts, which affects the smooth deployment of scene elements. Some implemented examples of text-to-scene conversion systems include the WordsEye system developed by ATT Labs (Ulinski et al., 2018) and the Text2scene system developed by the University of Virginia and IBM (Tan et al., 2019). However, these systems can only perform automatic conversion of text to static scenes. For sign language NLP, two types of systems may be required: one that converts natural language text to

simple animation generation, and another that creates interactive animated language. The former, which is still in its early stages and lacks a fully developed system, contrasts with the latter, which focuses on enabling users to control animated characters' actions and interactions through language. Representative systems include AnimNL (Badler et al., 2002) and Avatarclip (Hong et al., 2022), with AnimNL being applied to the ASL machine translation system. Additionally, text-to-scene conversion related to sign language also involves the issue of animation scripting, since this type of conversion requires interaction that allows the signer to control the layout and actions within the scene. The core task is to process the input language and map it to a specific point in the scene, where the signer can make purposeful modifications and settings through interaction.

This text-to-scene conversion highlights the importance of spatial concepts in human cognition and natural languages, including sign language. These concepts are essential for understanding relationships. The challenge in spoken language NLP is that spatial computing requires expertise in cognitive science and computer vision, making automatic spatial knowledge base creation difficult. Exploring text-to-scene conversion using sign language could advance spatial relationship understanding and virtual scene generation. On one hand, from the perspective of multi-channel representation in sign language, it considers content such as space, positions within space, and movement within space. On the other hand, it draws from sign language NLP to analyze spatial language semantics in spoken language. These two directions contribute to spatial information extraction, such as the spatial orientation and position of objects, combined with knowledge bases to eliminate the ambiguity in natural language, thus enabling the construction of 3D scenes.

4.2 Spatial Metaphor Computing

Spatial computing is closely related to spatial metaphors and involves the conversion between single-channel and multi-channel representations, while spatial metaphor refers to a type of conceptual metaphor where spatial relationships are used to understand and describe non-spatial concepts (Boroditsky and Lera, 2000). Sign language relies heavily on spatial metaphors, using space as both its conceptual framework and medium of expression. For instance, a Chinese deaf person can express success or failure with upward or downward gestures, exemplifying the conversion from single-channel to multi-channel representation. Similarly, spoken languages have vertical spatial metaphors, such as using "up" for the past and "down" for the future in East Asian languages. Thus, there's a correspondence between spatial metaphors in sign and spoken language (Gu et al., 2017). Understanding metaphors in spoken language focuses on developing models and algorithms for recognition and interpretation. However, it's unclear if these approaches apply to sign language metaphors, which possess both iconicity and metaphorical dual mapping, unlike spoken language's single mapping. Different cognitive agents, like deaf and hearing people, may interpret the same metaphor differently. Hence, statistical models alone are insufficient for metaphor comprehension; incorporating subjective knowledge and cognitive perspectives is necessary. Cognitive neuroscience methods, like ERP and fMRI, could explore how the brain processes sign language metaphors, leading to more scientific models for recognition and understanding of spatial metaphors.

Theoretical methods and models of sign language understanding need outer layer study, while brain mechanisms for comprehension require inner layer exploration. Only then can we seek an NLP foundation from human cognition and intelligence.

4.3 Classifier Predicate Computing

Since the 1960s, linguists have generally agreed that the linguistic phenomena of sign language can largely be explained through spoken language linguistics. However, classifier predicates in sign language represent a unique linguistic phenomenon (Cogill-Koez and Dorothea, 2000). This means that if the currently popular Gloss annotation is used as a single-channel encoding for sign language, all other grammatical phenomena can be addressed using NLP, except for classifier predicates. Deaf signers encounter a classifier predicate almost every minute during communication, with certain types appearing as many as 17 times (Morford and MacFarlane, 2003). Since classifier predicates are the most complex phenomenon in sign language NLP (often related to spatial semantics), they challenge traditional definitions of language expression. The key issue is how to represent the handshapes and motion types of classifier predicates and how to map them to semantic representations. To compute classifier predicates, map objects to mental and

physical spaces. During encoding, spatial information must be quantified as morphemes. Multiple morphemes often express a classifier predicate's full meaning. The spatial information conveyed is more complex than imagined, especially when describing relationships between objects, where it becomes intricate. Liddell (2003) conducted a statistical analysis of the simple classifier predicate "one person walking toward another person" and found that 28 morphemes were needed to fully express the spatial information. This spatial information includes: the two people facing each other, a specific distance between them, movement along a straight path, being on the same horizontal plane, and standing in a vertical alignment, among others. Based on this, Liddell evaluated classifier predicates as a non-spatial, multi-morpheme structural model, because in order to fully express the various spatial information, the number of morphemes required for a classifier predicate can be vast, potentially even infinite.

Since the expression of classifier predicates is dynamic, it is also necessary to encode the interactions between entities and the constraints of the 3D scene deployment into a series of rules(Bangham et al., 2000). To effectively convey a scene like "one person walking toward another," it's essential to establish the entities' positions and choose the starting and ending points for the moving entity. The signer must decide if the path is linear or curved and whether it's bumpy or smooth. Additionally, while expressing classifier predicates, the road and ground plane are communicated. To avoid errors like depicting someone moving below the ground, basic common and world knowledge is required, such as the understanding that people typically stand on the ground plane. Clearly, the use of classifier heavily depends on semantic understanding, spatial knowledge, and logical reasoning.

The two difficulties mentioned above contribute to the complexity of classifier predicate computing, to the point where sign language scholars have described it as supra-linguistic spatial sign language and as constituting spatial parametric expressions (Wehrmeyer, 2022). Currently, aside from the ZARDOZ system (Veale et al., 1998), no other sign language machine translation systems have managed to solve classifier predicate computing. Classifier predicates are uniquely complex in linguistic computation and could be a key feature in sign language NLP. They involve converting and mapping from one channel to multiple channels, with a significant use of spatial metaphors and scene-processing. Research should focus on brain processing of these predicates and develop small systems to simulate intelligent behavior. With sufficient understanding and clarification of cognitive mechanisms, a comprehensive solution for classifier predicate computing can be developed in the future.

In summary, addressing these theoretical challenges—text-to-scene conversion, spatial metaphor computing, and classifier predicate computing—requires moving beyond single-channel, text-based NLP approaches toward methods that fully capture sign language's multi-channel and spatial nature. We propose three foundational directions to facilitate this shift: (1) scalable tools for automatic multi-channel tagging, reducing reliance on manual annotation by systematically capturing handshape, movement, facial cues, and other channels; (2) neural models integrating spatial reasoning, enabling learning architectures to incorporate 3D layout, trajectory, and scene constraints inherent in sign language; and (3) community-driven data collection, ensuring that large-scale, culturally diverse corpora accurately reflect real-world signing practices.

5 Conclusion

Sign language NLP remains challenging due to multi-channel features and low-resource constraints exacerbated by regional variations. The single-channel paradigm of spoken language NLP fundamentally conflicts with sign language's spatial-temporal dynamics, requiring precise synchronization that current models lack. While multi-channel architectures show promise, they struggle to capture structural complexity without linguistically grounded annotation. To overcome these hurdles, it is essential to focus on a deeper understanding of the foundational structure of sign language itself. Advancing NLP for sign languages will depend on grounding these approaches in the intrinsic features of sign language, ensuring that future developments are both linguistically accurate and contextually sensitive.

6 Limitations

6.1 Lack of Large-Scale Empirical Validation

This paper primarily presents a theoretical and comparative analysis of sign and spoken language processing. While we highlight key computational

challenges and propose potential directions, our work does not include large-scale empirical experiments to validate these claims. Future research should conduct systematic evaluations using real-world sign language corpora. This will help assess the effectiveness and feasibility of the proposed approaches in addressing the identified challenges.

6.2 Data Scarcity and Annotation Bottlenecks

A fundamental challenge in SLP is the low-resource nature of sign language data. While we discuss the need for scalable multi-channel annotation tools, this remains an unsolved issue. Current datasets are often limited in size, linguistic diversity, and coverage of sign languages worldwide. Moreover, manual annotation remains costly, and the lack of standardized multi-channel representations further complicates dataset expansion.

6.3 Computational Challenges in Multi-Channel Processing

Sign language's multi-channel feature—combining handshape, movement, facial expressions, and body posture—poses significant computational challenges. Existing multimodal models often focus on fusing speech, text, and vision but lack fine-grained mechanisms for capturing interdependent linguistic features in sign language. While we suggest spatial reasoning and 3D-aware architectures, their practical effectiveness remains untested on large sign language corpora.

6.4 Theoretical Assumptions and Model Adaptability

Our discussion is based on linguistic and cognitive insights into sign language structure, which may not fully align with the computational constraints of NLP systems. While we advocate for moving beyond single-channel NLP paradigms, transitioning toward multi-channel, spatial-semiotic models requires substantial architectural modifications, which could present scalability and efficiency concerns.

6.5 Generalizability Across Different Sign Languages

This paper mainly considers general linguistic properties of sign languages but does not explicitly address cross-linguistic variations. Different sign languages (e.g., ASL, BSL, CSL) exhibit unique syntactic structures, lexical variations, and cultural adaptations, which may impact the applicability of proposed computational frameworks. A one-size-fits-all model for sign language processing remains a challenge.

6.6 Ethical Considerations and Community Involvement

While we emphasize the need for community-driven data collection, ethical concerns such as informed consent, representation of diverse Deaf communities, and accessibility in AI-driven sign language applications require further discussion. Engaging Deaf researchers and sign language users in the development of computational models is essential to ensure inclusivity and fairness in SLP technologies.

References

Al Abdullah, Ghada Amoudi, and Hanan Alghamdi. 2024. Advancements in sign language recognition: A comprehensive review and future prospects. *IEEE Access*.

Aronoff, Markand, Janie ReesMiller, and eds. 2020. *The Handbook of Linguistics*. John Wiley & Sons.

Norman I. Badler, Jan M. Allbeck, Liwei Zhao, and Meeran Byun. 2002. Representing and parameterizing agent behaviors. *Proceedings of Computer Animation 2002 (CA 2002)*, pages 133–143.

Bangham, J. Andrew, and et al. 2000. Virtual signing: Capture, animation, storage and transmission-an overview of the visicast project. In *IEE Seminar on speech and language processing for disabled and elderly people (Ref. No. 2000/025)*. IET.

Boroditsky and Lera. 2000. Metaphoric structuring: Understanding time through spatial metaphors. *Cognition*, 75(1):1–28.

Bragg, Danielle, and et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*.

Diane Brentari. 2019. *Sign Language Phonology*. Cambridge University Press.

Cassidy, Steve, and et al. 2018. Signbank: Software to support web based dictionaries of sign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Cogill-Koez and Dorothea. 2000. Signed language classifier predicates: Linguistic structures or schematic visual representation? *Sign Language & Linguistics*, 3(2):153–207.

da Rocha Costa, Antonio Carlos, and Graçaliz Pereira Dimuro. 2003. Signwriting and swml: Paving the way to sign language processing. In *Atelier Traitement Automatique des Langues des Signes, TALN*.

Dreuw, Philippe, and et al. 2008. Benchmark databases for video-based automatic sign language recognition. In *Proceedings of Language Resources and Evaluation Conference (LREC)*.

Shannon C E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

ELAN. 2024. (version 6.9) [computer software] max planck institute for psycholinguistics, the language archive. https://archive.mpi.nl/tla/elan.

Karen Emmorey. 2001. *Language, cognition, and the brain: Insights from sign language research*. Psychology Press.

Karen Emmorey, Rain Bosworth, and Tanya Kraljic. 2009. Visual feedback and self-monitoring of sign language. *Journal of Memory and Language*, 61(3):398–411.

Karen Emmorey and David Corina. 1990. Lexical recognition in sign language: Effects of phonetic structure and morphology. *Perceptual and motor skills*, 71(3_suppl):1227–1252.

Fei, Hao, and et al. 2024. From multimodal llm to human-level ai: Modality, instruction, reasoning, efficiency and beyond. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*.

Ione Fine et al. 2005. Comparing the effects of auditory deprivation and sign language within the auditory and visual cortex. *Journal of Cognitive Neuroscience*, 17(10):1621–1637.

Yan Gu et al. 2017. Conceptual and lexical effects on gestures: the case of vertical spatial metaphors for time in chinese. *Language, Cognition and Neuroscience*, 32(8):1048–1063.

Hanke and Thomas. 2002. ilex - a tool for sign language lexicography and corpus analysis. In *LREC*.

Hanke, Thomas, and et al. 2020. Extending the public dgs corpus in size and depth. In *Proceedings of the LREC2020 9th workshop on the representation and processing of sign languages: Sign language resources in the service of the language community, technological challenges and application perspectives.*

Hickok, Gregory, and et al. 1996. The basis of hemispheric asymmetries for language and spatial cognition: Clues from focal brain damage in two deaf native signers. *Aphasiology*, 10(6):577–591.

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*.

Yujing Huang and Fernanda Ferreira. 2021. What causes lingering misinterpretations of garden-path sentences: Incorrect syntactic representations or fallible memory processes? *Journal of Memory and Language*, 121:104288.

Matt. Huenerfauth. 2004. Spatial representation of classifier predicates for machine translation into american sign language. In *Workshop on Representation and Processing of Sign Language, 4th Internationnal Conference on Language Ressources and Evaluation (LREC)*.

Matt Huenerfauth. 2005. American sign language generation: multimodal nlg with multiple linguistic channels. In *Proceedings of the ACL Student Research Workshop*.

M. Kopf, M. Schulder, and T. Hanke. 2022. The sign language dataset compendium: Creating an overview of digital linguistic resources. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 102–109.

Deyi Li. 2024. Cognitive computing in the era of big data. https://tech.china.com/news/mobnet/11103682/20131113/18144680.html.

Scott K. Liddell. 2003. Sources of meaning in asl classifier predicates. In *Perspectives on classifier constructions in sign languages.*, pages 209–230. Psychology Press.

Albert Mehrabian. 1973. Linguistics: Silent messages. *American Anthropologist*, 75(6):1926–1927.

Marcelo A. Montemurro and Damián H. Zanette. 2001. Entropic analysis of the role of words in literary texts. *Adv. Complex Syst.*, 5:7–18.

Jill P. Morford and Jill Patterson James MacFarlane. 2003. Frequency characteristics of american sign language. *Sign Language Studies*, 3:213 – 225.

Carol Neidle, Augustine Opoku, Gregory Dimitriadis, and Dimitris Metaxas. 2018. New shared & interconnected asl resources: Signstream® 3 software; dai 2 for web access to linguistically annotated video corpora; and a sign bank. In *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, LREC*, pages 147–154, Miyazaki, Japan.

Carol Neidle and Christian Vogler. 2012. A new web interface to facilitate access to corpora: Development of the asllrp data access interface. In *The 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC*, Istanbul, Turkey.

Wendy Sandler. 2012. The phonological organization of sign languages. *Language and Linguistics Compass*, 6(3):162–182.

Rachel Sutton-Spence and Bencie Woll. 1999. *The linguistics of British Sign Language: an introduction*. Cambridge University Press.

Fuwen Tan, Song Feng, and Vicente Ordonez. 2019. Text2scene: Generating compositional scenes from textual descriptions.". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Ulinski, Morgan, Bob Coyne, and Julia Hirschberg. 2018. Evaluating the wordseye text-to-scene system: imaginative and realistic sentences. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Clayton. Valli. 2000. *Linguistics of American Sign Language: An Introduction*. Clerc Books.

Veale, Tony, Alan Conway, and Bróna Collins. 1998. The challenges of cross-modal translation: English-to-sign-language translation in the zardoz system. *Machine Translation*, 13:81–106.

Ella Wehrmeyer. 2022. Psycholinguistic errors in signed simultaneous interpreting. *Interpreting*, 24(2):192–220.

Dengfeng Yao, Jiang Minghu, Hong Bao, Hanjing Li, and Abulizi Abudoukelimu. 2019. Thirty years beyond sign language computing: Retrospect and prospect (in chinese). *Chinese Journal of Computer*, 42(1):1–28. (EI).

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.