Digital tools for sign language research: towards recognition and comparison of lexical signs

Fragkiadakis, M.

CHAPTER 1

Introduction

It is a prevalent fallacy that all deaf people, throughout the world, understand the same sign language. This, however, is not the case. Sign languages are particular to specific groups and cultures, and they emerge spontaneously as a result of interactions amongst deaf people within such communities. They are not created by hearing people or based on spoken languages. The grammatical structure, lexicon, and cultural background of each sign language are unique. Different sign languages are used in different nations, such as American Sign Language, British Sign Language, and Chinese Sign Language, each having their own specific features. Furthermore, geographical differences within a nation and dialects within a sign language exist. This linguistic diversity among the deaf community is a reflection of the unique ways in which they communicate.

Over the years, Deaf communities have campaigned for the recognition of signed languages as legitimate languages as well as the right to study and use them [178]. Signed languages are sophisticated modes of communication that are at least as functional linguistically and socially as spoken languages. Deaf people are, nevertheless, frequently encouraged to use spoken languages through lipreading or text-based communication. Suppression of signing in favor of spoken languages is exacerbated by the absence of signed languages from current language

technologies and tools [178].

Like all natural languages, signed languages encompass phonological, morphological, syntactic, and semantic structural layers and fulfill identical social, cognitive, and communicative roles. However, the difference lies in the mode of communication. While oral-auditory channels form the basis of spoken languages, signed languages mainly rely on the visual-gestural system. This visual-gestural approach employs the signer's facial expressions, hands and body movements, and the space around them to differentiate and convey meanings [119].

The study of sign languages has long been an important area of research for linguists and educators. In addition to learning about the structure and use of sign language, it is also important to understand the culture and history of the deaf community and the role that sign language plays in different contexts.

Over the years, technology has been instrumental in the archiving, examination, and research of sign languages. Utilizing video technology, sign language content can be recorded and disseminated for pedagogical and investigative objectives. Furthermore, technology has facilitated the creation of digital dictionaries and various learning aids for sign languages, and has served as a critical tool for monitoring and assessing the progression and evolution of sign languages over time. Finally, technology has been employed to develop applications and software for sign language users, such as translation apps and virtual environments for practicing and learning the languages in question.

In this thesis, we look more closely at how technology has influenced the study and usage of sign language. We examine the different tools and resources that have been developed, as well as the potential advantages and disadvantages of these technologies. Through this investigation, we seek to get a deeper understanding of the role that technology plays in the ongoing research and development of sign languages. To that end, we have not merely surveyed the field but contributed to it by developing novel tools designed to facilitate research and enrich the understanding of sign languages. These tools, as detailed in the subsequent sections, are primarily concerned with automatic annotation of sign and gestural sequences, detection of handshapes and handedness, and enhancing search capabilities within sign language dictionaries. By doing so, we aim to reinforce the symbiotic relationship between technology and the ongoing study and growth of sign languages.

1.1 Sign Language phonology and annotation

In this section we explore the structural intricacies and analysis tools essential to sign language research. Starting with an investigation into sign languages' fundamental elements in the "Sign Language Phonology" section, we shift towards strategies for their documentation and analysis in "Sign Language Annotation."

1.1.1 Sign language phonology

As Stokoe discusses [152], signs are comprised of several components or articulators. These articulators include non-manual components, such as head movement and torso posture, as well as manual components including hand configuration, palm orientation, placement, contact, and local movement.

In his seminal research [152, 153], Stokoe developed a notation system, now referred to as "Stokoe notation," that breaks down each American Sign Language (ASL) sign into three phonological elements: handshape, position (articulation site), and movement. In conjunction with the direction the palm of the hand is facing, these elements are intrinsic phonological factors necessary for a thorough and effective characterization of ASL signs. Alterations in any of these components could potentially lead to the formation of a distinct sign, thereby changing the meaning of the sign [152].

The distinct traits and motions of the hands and fingers when producing signs are referred to as "manual elements." The form and arrangement of the fingers and hands are referred to as hand configuration, whilst the direction the palm is facing is referred to as palm orientation. Placement refers to where the hands are in respect to the body or the signing area, and touch all physical interactions that take place during signing.

Path movement refers to the path taken by the hands or fingers as they travel across space, whereas local movement refers to the more minute motion made by certain handshape or finger configurations during a sign realization. These manual elements work together to create the visual representation of the sign.

Non-manual components are also very important in sign language. Head movement like nodding or tilting can convey grammatical and semantic information. Torso posture refers to the positioning of the body,

including the shoulders, chest, and back, which can contribute to the overall expression and meaning of a sign.

1.1.2 Sign language annotation

In contrast to spoken languages, there is no standardized technique for writing signs comparable to the International Phonetic Alphabet (IPA). Attempts to develop a sign writing system, such as SignWriting or HamNoSys, have been undertaken, however these systems are not utilized consistently across various studies. A single sign can include a large amount of phonological information, such as handshape, movement, position, and orientation, and all of these aspects, as well as their interaction, must be represented by distinct symbols. An example of how a sign can be annotated using different systems is presented in figure 1.1.

The absence of a universally accepted orthography in sign languages presents a considerable obstacle to cross-linguistic comparison and the creation of sign language corpora. This issue primarily manifests in two interconnected challenges: the scarcity of resources and the inconsistency in data formats.

With regards to resource scarcity, the available data for machine learning applications is limited, which hampers the advancement of sign language recognition and translation tools. Collecting and processing new data is a resource-intensive endeavor, further exacerbating the problem.

The second challenge revolves around the variation in data formats. Annotation formats, which can be viewed as a form of orthography, are not standardized and differ greatly among various resources. The current data formats often prove unsuitable for machine learning applications, impeding the development of neural model-based automatic tools.

In essence, the lack of a shared orthography in sign languages not only hinders the progression of sign language recognition and translation technologies but also complicates the process of cross-linguistic comparison in sign language research [44].

HamNoSys - iLex

The Hamburg Notation System for Sign Languages (HamNoSys) "is an alphabetic system that describes signs on a mostly phonetic level", as

described by T. Hanke [74]. The system was first introduced in 1984 and has undergone significant development over the years, with the current version being 4. The primary goals of the system, as outlined by Hanke [74], are to be internationally usable, iconic, economical, integrate with standard computer tools, have a formal syntax, and be extensible.

A significant advantage of HamNoSys, compared to other sign language notation systems, is its commitment to iconicity, aiming to capture the visual and expressive qualities inherent in sign languages. The system utilizes symbols and annotations to represent the physical movements, handshapes, and facial expressions involved in signing. This iconic representation enhances the accuracy and comprehensibility of sign language descriptions. For example, a sign depicting a flying bird in a specific sign language can be visually conveyed through appropriate HamNoSys symbols and annotations, ensuring that the iconic nature of the sign is preserved and universally understood.

HamNoSys also emphasizes economy in sign language description, both in terms of simplicity and concise representation. The system employs a formal syntax that allows users to represent complex signs using a compact set of symbols and rules. This efficiency is particularly valuable for constructing sign language dictionaries, educational resources, and research databases where a large number of signs need to be organized and described systematically. HamNoSys enables a succinct representation of signs, saving time and effort for sign language researchers and educators.

However, HamNoSys faces certain challenges. One significant challenge is the learning curve associated with mastering the system's notation. HamNoSys employs a specialized set of symbols and rules, requiring substantial time and effort to become proficient. Novice users may initially struggle with accurate interpretation and the composition of complex sign descriptions. This learning barrier could hinder the widespread adoption of HamNoSys, especially among sign language users who have limited exposure to linguistic notation systems.

In complement to HamNoSys, the development of the iLex system offers additional benefits. iLex, or "integrated lexicon," serves as a transcription database for sign languages and a lexical database. It complements HamNoSys by providing an extensive collection of sign language transcriptions, enhancing the accessibility and availability of sign language data for researchers and educators.

ID glosses - ELAN

iLex is an incredibly robust and content-dense system, however, it lacks the immediacy and intuitive usability that researchers often need. Its sophisticated structure and functionalities necessitate a certain level of investment in time and effort to master. In light of these challenges, researchers have sought more efficient methodologies to generate machine-readable corpora.

One such approach involves the use of uniquely identifying spoken language words, which typically refer to a specific sign, serving as unique identifiers or "ID-glosses" [88]. This technique facilitates the annotation and search processes by providing clear labels for individual signs, thereby streamlining the entire procedure.

The ID-gloss essentially acts as a bridge between the signed and spoken languages, providing a way for researchers to refer to specific signs using spoken language terms. This is particularly useful when working with machine-readable corpora, as it allows for the efficient indexing, searching, and processing of sign language data. However, it's important to note that the use of ID-glosses isn't a direct translation of signs into spoken language words. Rather, they are a form that researchers use to refer to specific signs.

However, it is important to note, as pointed out by Johnston [87], that the application of ID-glosses is contingent on the existence of a reference lexical database, a comprehensive dictionary that is a product of foundational research into the lexicon. These databases provide the necessary ID-glosses during the annotation process, acting as invaluable language resource sites. Thus, these repositories of lexical information are instrumental in effectively utilizing ID-glosses in research efforts [87].

By leveraging such databases, researchers can significantly enhance the efficiency and accuracy of their work, making advancements in the field of linguistic research more attainable. Notwithstanding, it underscores the importance of foundational research and the creation of comprehensive lexical databases to support future linguistic endeavors.

A system that is commonly used for manual annotation with ID glosses is ELAN [149]. Developed at the MPI for Psycholinguistics, ELAN is a Java and XML based annotation tool that has been widely used in a number of sign language corpus projects [40, 143]. Multiple layers (i.e. tiers) can be used for the annotations and they can be interdependent hierarchically.

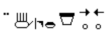
	SignWriting Printing	SignWriting Shorthand	Stokoe Notation	HamNoSys Notation
what? 				

Figure 1.1: Representation of the ASL sign WHAT in three notation systems [146].

1.2 Sign Language Processing

Having delved into the complexity of sign language's structures and annotation techniques, we now transition into an innovative domain, "Sign Language Processing." This section will illuminate how our existing knowledge intersects with artificial intelligence and computer vision, introducing a compelling new frontier in the field of sign language studies.

Sign language processing (SLP), as recognised by Bragg et al. [19] and Yin et al. [178], is a growing research field derived by the interception of natural language processing (NLP) and Computer Vision (CV). NLP refers to the emerging field of artificial intelligence that is concerned with developing all the necessary tools to allow computers to understand text and spoken words in a similar fashion that humans do. Similarly, computer vision enables computational systems to extract meaningful information from visual inputs such as videos and images. Thus, SLP is often concerned with challenges such as the recognition of sign language from video material (sign language recognition), the machine translation to spoken language text (sign language translation) or the translation of spoken language into sign language (sign language production) [119].

1.2.1 Sign Language Recognition

Sign language recognition is the task of recognising discrete signs as glosses from video material [4]. Early attempts focused on the identification of isolated glosses by making use of hand crafted techniques and features while frequently utilising special recording devices such as

microsoft's Kinect [167] and power gloves [91]. To model the changes of the predicted spatial features in time (temporal modeling) studies often employed classical sequence learning models, such as Hidden Markov Models (HMMs) [51, 181, 98] and hidden conditional random fields [169].

The emergence of deep learning networks and techniques provided a noteworthy improvement in overall performance and accuracy of models developed for video-related tasks such as human action recognition [45, 54] and gesture recognition [117, 26]. As a result, sign language recognition shifted towards the use of such networks and techniques as well.

Deep learning is a sub field of artificial intelligence that uses algorithms to imitate the operations of the human brain in order to enable machines to learn and ultimately predict specific outcomes. It is based on the notion of neural networks, which are networks of linked nodes that process information in the same way as neurons in the human brain do. Deep learning techniques are used to discover patterns and anticipate outcomes in a variety of applications such as natural language processing, computer vision, image recognition, and robotics.

Cui, Liu, and Zhang [42] developed a weakly supervised framework for continuous sign language recognition by encoding the video using a spatio-temporal Convolutional Neural Network encoder and predicting a gloss using a Connectionist Temporal Classification (CTC). Subsequently, they encode each gloss-level segment separately, trained to predict the gloss category, and use the encoding of gloss video segments to improve the sequence learning model.

Conversely, Cihan Camgöz et al. [23] take a different approach and formulate the problem as a natural-language translation problem. They use AlexNet convolutional neural network [100] to encode each video frame. The gloss is then generated using a GRU encoder-decoder architecture with Luong Attention [109]. In their subsequent study, Camgöz et al. [24] substituted the GRU with a transformer encoder [164], and employed a Connectionist Temporal Classification (CTC) for gloss decoding. This approach, they demonstrated, resulted in a modest enhancement for the video-to-gloss task.

More recently, Adaloglou et al. [4] evaluated different computer vision-based methods for the video-to-gloss challenge in a comparative experimental study. The researchers scrutinized different methodologies from past studies [25, 41, 90] across several datasets [79, 23, 166, 90] for both isolated and continuous sign language recognition. Their argument

posits that 3D convolutional models deliver superior performance compared to models solely reliant on recurrent networks and that these models exhibit a higher degree of scalability.

Finally, Zuo and Mak [183] proposed new techniques to enhance the training of deep learning-based continuous sign language recognition (CSLR) models, with a particular focus on improving the visual and sequential components of these systems. Recognizing that the visual aspect of sign language interpretation often suffers from insufficient training, the authors propose an auxiliary constraint that inserts a keypoint-guided spatial attention module into the visual component. This makes the system better at focusing on critical areas such as the signer's face and hands, improving spatial attention consistency.

1.2.2 Sign Language Translation

The task of translating sign language from video to text in a spoken language is commonly referred to as sign language translation. The majority of previous research on Sign Language Recognition has tackled the challenge as a simple gesture identification problem, neglecting the linguistic aspects of the sign language and presuming a one-to-one mapping of sign to spoken words [23]. Fundamentally, translating sign videos to spoken language is a sequence-to-sequence learning challenge. It seeks to understand the signs' spatio-temporal representation, the relationship between them (the language model), and how these signs map to a spoken or written language [23].

A common approach to Sign Language Translation is to consider the challenge as a two-fold task also known as Gloss-to-Text. First a Sign Language Recognition (SLR) system is used to extract the sign language glosses from videos after which a translation system (SLT) generates spoken language translations from these glosses.

Due to the extensive application of transformer-based deep learning models in spoken language processing, Yin and Read [177] introduced a transformer encoder-decoder model [164]. Their model demonstrated advantages on both the RWTH-PHOENIX-Weather-2014T (DGS) and ASLG-PC12 (ASL) datasets. They showed that their model performing video-to-text translation outperformed the translation of ground truth glosses challenging the use of glosses as a representation of sign language.

Another part of sign language translation is the the task of translation between a spoken language text and sign language glosses

known as text-to-gloss [119]. For translating between English sentences and American Sign Language glosses, Zhao et al. [182] developed a Tree Adjoining Grammar (TAG)-based approach. A Tree-Adjoining Grammar (TAG) is a tree rewriting system [89] using basic trees as its primitive constituents. These trees are anchored by lexical elements such as nouns and verbs in a lexicalised TAG [145]. The subjects and objects of verbs, adjectives, and other predicates have argument locations in the elementary trees, which limit how they may be joined and establish the predicate-argument structure of the input phrase.

The full process of converting a raw video containing sign language to spoken language text is known as video-to-text. Camgöz et al. [24] presented a unique transformer-based architecture that learns Continuous Sign Language Recognition and Translation simultaneously while being trainable end-to-end. They encode each frame separately using the pre-trained spatial embeddings from Koller et al. [99] and utilize a transformer to encode the frames. The classification of the sign language gloss is held by the use of a Connectionist Temporal Classification (CTC) method [68]. They subsequently employ a transformer decoder to decode the spoken language text, one token at a time, using the same encoding.

Following up, Camgöz et al. [22] present a novel design that eliminates the need for gloss supervision and incorporates both manual and non-manual features into the Sign Language Translation task. To collect three different data channels by isolating the signing hand and face while also performing 3D pose estimation. They investigate many methods for combining the data in these channels, including early and late fusion in the transformer architecture. Experiments on the RWTH-PHOENIX-Weather dataset show that their technique is comparable with the current state-of-the-art while also limiting the need on glossed-annotated data.

Using a pose estimating framework on the raw video data and then converting the predicted pose sequence to spoken language text is one way to lessen the complexity of the recognition aspect of the sign language translation challenge. By feeding the pose sequences into a translation model based on a sequence-to-sequence architecture, Ko et al. [96] experimented with different attention mechanisms and reported a high translation accuracy.

1.2.3 Sign Language Production

The technique of creating a sign language video from spoken language text is known as sign language production. As Rastgoo et al. discuss [135], sign language recognition and production are two necessary parts for making a "robust system capable of translating the spoken languages into sign languages and vice versa". In their review study they differentiate between five categories of approaches in Sign Language Production (SLP).

Avatars, a common approach in SLP, are 3D animated models that produce movements with the fingers, hands, facial expressions, and body and have been developed to work in a variety of sign languages [135]. Nevertheless, the avatar approaches face issues such as under-articulation, awkward and unnatural motions, and the absence of non-manual information such as facial expressions. To overcome some of these issues and to produce more realistic results, researchers have used the data collected from motion capture devices. However, as Rastgoo et al. [135] point out, due to the high cost of the collection and annotation process of such data, the results of these studies are limited to a small set of phrases.

Stoll et al. [154] presented a hybrid approach combining Neural Machine Translation (NMT), Generative Adversarial Networks (GANs), and motion creation for automatic SLP. A major advantage of their method is the minimum need for gloss and skeleton level annotations. The suggested approach starts by first translating spoken language sentences into sign pose sequences. Then, in order to produce plausible sign language video sequences, a generative model is utilized.

However, as Rastgoo et al. indicate [135], while NMT-based systems have shown to be effective in translation tasks, there are still some significant obstacles to overcome. First and foremost, domain adaptation is a critical criterion in designing machine translation systems customized to a given use case. The second issue is the amount of training data available. Increasing the amount of data in deep learning-based models, in particular, can improve results. The last challenge is the treatment of unknown words as all translation models perform poorly on them.

Another approach to SLP is the use of a directed graph developed from motion capture data mostly known as a Motion Graph (MG). MG is a computer graphic method for dynamically animating figures

that can create new sequences to achieve certain objectives [135]. MG is often paired with an NMT-based network that can be used for continuous-text-to-pose translation. MG confronts various obstacles, despite the fact that it can produce realistic and controlled motion. The first issue to be taken into account is the overall need for large data-sets in order to demonstrate the model's capability with a really diversified variety of behaviors. Other issues in MG development are the graph's scalability and computational difficulty in selecting the appropriate transitions [135].

With recent breakthroughs in deep learning, new techniques using neural network-based architectures, such as Convolutional Neural Networks (CNNs) [33, 161], Recurrent Neural Networks (RNNs) [70, 160], Variational AutoEncoders (VAEs) [94, 174], and Generative adversarial networks (GANSs) have been used in the field of automatic image and video generation and synthesis. Similarly to sign language translation, different sub-domains have been emerged with various representations as starting and end points.

In the field of robotics and animation, pose-to-video, also known as motion-transfer or skeleton animation, is the translation of a series of poses into a realistic-looking video. This process has been utilised by sign language production studies as well.

Ventura et al. [165] investigated if and how effectively Deaf people can interpret automatically produced sign language videos. Using the state-of-the-art human motion transfer technique from the "Everybody Dance Now" study [31], they created realistic videos. They first extracted keypoints from the raw video by employing OpenPose [27]. Using the model from Wang et al. [170], the keypoints were then utilized to condition a Generative Adversarial Network (GAN) to create each video frame. They quantitatively and qualitatively assessed the produced videos, demonstrating that existing models are unable to generate suitable videos for Sign Language due to a lack of detail in the hands.

Later on, Saunders et al. [141] proposed an SLP model for photo-realistic continuous sign language video synthesis directly from spoken language. To manage the translation from the spoken language to the skeletal poses, they adopt a transformer architecture with a Mixture Density Network (MDN) formulation. The skeletal pose sequence is then sent into a pose-conditioned human synthesis model, which produces a photo-realistic sign language video. They also present a new keypoint-based loss function to increase the quality of the hand

synthesis as previously identified by Ventura et al. [165] as the factor that impacts SLP performance.

1.3 Challenges in Sign Language technologies

Machine learning and deep learning can help address some of the main challenges for developing technologies for sign language processing and translation by offering a more comprehensive and accurate set of tools for recognizing, interpreting, and translating sign language. Machine learning algorithms can be used to learn the intricate patterns of sign language, while deep learning algorithms can be used to build more sophisticated and accurate models of recognizing sign language. With these tools, developers can create systems that are able to accurately recognize, interpret, and translate sign language in real-time. Additionally, deep learning algorithms can help to create more natural and accurate translations of sign language into other sign or spoken languages. Nonetheless, present techniques are limited in a variety of ways, making them often unsuitable for use in current sign language research. In this subsection we discuss some of main limitations these techniques pose.

Firstly, the training of deep neural networks often requires a substantial amount of data and robust computer capacity. This data typically consists of manually annotated video material. Consequently, under-documented sign languages that lack extensive manual annotations are often neglected as training material. The consequence of this neglect is the limited development of automatic sign language recognition or translation tools for these languages, creating a substantial barrier for many Deaf and hard-of-hearing individuals who rely on these languages for communication.

Furthermore, a general issue with these deep neural networks is their lack of interpretability. While these models might successfully recognize or translate a particular sign language, it's difficult to understand why they made a certain decision or prediction. This can create challenges in refining the system's performance, as it's harder to identify and address the specific areas where the model might be going wrong. This is not necessarily the case with machine learning models, whose output can be more explicable. A subsequent effect from the lack of transparency is that users might have hard time on trusting the systems' output, which

is especially problematic in scenarios where accuracy is critical.

In addition, these networks are typically trained in a single sign language and struggle to perform optimally in others due to each sign language's unique complexities. However, it's important to note that recent studies have explored the use of "transfer learning", a technique that employs an already trained neural network—usually in a sign language with large annotated video materials—for a different context or language [15, 53, 34]. Despite the promise of this technique, significant challenges remain, especially in relation to the standardization and volume of data available for under-documented sign languages, and the overarching issue of interpretability.

Moreover, most well annotated sign language data sets used in machine and deep learning applications have been filmed in controlled environments with sufficient lighting and position of the signers in the frame. As a result, it is unclear to what degree these systems, evaluated in such data sets, can be reproduced in corpora filmed in real-world conditions. It is common for sign language and gesture corpora, particularly those recorded outside of studio settings, to be of poor quality or low resolution, dimly illuminated, and to frequently have many people in the frame (an example of such conditions is shown in figure 1.2). These attributes create an additional challenge to the evaluation process of the accuracy of such applications and often are omitted for the training process.

For those learning sign languages, as well as educators, parents, and scholars who are researching the languages in question, online video based sign language lexica are an invaluable resource. Fundamentally, these lexica let a user submit a query with the gloss of the requested sign and receive a video or a picture of that sign. Along with this capability, certain lexica allow the user to specify some of the formal parameters of the target sign, such as its position, handshape, or movement, and obtain all the signs that include these properties.

Despite being a user-friendly feature in sign language lexica, a sign search functionality based on formal parameters still requires dictionary compilers - the people who collect, organize, and describe the sign language data - to manually connect these values to the various sign videos. This manual linking process is similar to the process of annotating sign language corpora, which involves adding explanatory or interpretative information to the sign language data. This process is a rather complex and time-consuming task. It often involves watching

Figure 1.2: Frame from Dogon Sign Language corpus and performance of OpenPose pose estimation framework.

videos of signs multiple times and meticulously documenting each parameter. Due to the high amount of effort and time required, few dictionaries currently offer a search function based on formal parameters.

Furthermore, another challenge is the possibility of human error in the process of linking formal parameters to signs. Mistakes can lead to inaccurate search results, which can impact the utility and reliability of the dictionary.

Sign language lexica can also be a valuable source for historical-comparative research on sign languages. This type of research frequently employs methods such as lexical comparison and lexicostatistics [172, 86, 13, 113]. Lexicostatistics is a branch of linguistics that uses statistical methods to compare lexical items in different languages to determine their degree of similarity and, subsequently, the potential common ancestral roots they may share.

More specifically, lexicostatistics involves measuring the degree of similarity in parameter values across different sign languages. Studies have investigated the extent of lexical overlap or similarity on a pre-determined list of terms [18]. By doing so, they can estimate the likelihood that two sign languages share a common lineage.

This means that researchers take specific words or signs, analyze their formal parameters like handshape, movement, and location, and then statistically compare these to find any commonalities [18]. This method can help illuminate how sign languages have evolved and branched off from one another, providing deeper insights into the history and development of sign languages worldwide.

The application of lexicostatistics in sign languages is not without its challenges. One of the main issues is the lack of standardization in the similarity criteria used in studies. Parks [128] points out that many studies do not use a consistent set of similarity criteria, resulting in different outcomes for each comparison due to the use of various standards. This can make it difficult to compare results across different studies and to draw meaningful conclusions.

Another issue, as discussed before, is that there is no formal transcription technique for sign languages. This implies that different studies employed various annotation methods to encode the sign form parameters. This lack of uniformity in annotation methods can also make comparing results between different studies challenging. The usage of diverse annotation methods might cause comparability issues, which can reduce the accuracy of lexicostatistics findings.

In summary, while lexicostatistics can be a valuable tool for historical-comparative research on sign languages, the lack of standardization in similarity criteria and annotation systems, as well as the complexity of sign languages, can make it challenging to apply.

1.4 Research Questions

This dissertation aims to delve deep into the heart of critical challenges associated with the application of machine and deep learning techniques to sign language processing, recognition and comparison as outlined in the previous section. Our intention is to break down these complexities and explore potential solutions through our research. Consequently, our investigation in this dissertation is steered by the following principal questions that evolved during this research:

1. How effective are machine and deep learning methodologies in processing and recognizing sign languages?

2. How can machine learning methodologies be employed to develop a system that automatically predicts and annotates sign and gestural sequences from video material, and what is the effect of this automation on the efficiency of sign language research?

3. How can a reverse search tool be engineered to allow users to sign a query and retrieve matching signs from a sign language dictionary, and how does the proficiency level of the user in a sign language influence the performance of this tool?

4. How do different computational methodologies, such as Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP), Dynamic Time Warping (DTW), and Neural Networks, perform in a reverse sign language search tool, particularly in terms of changes in performance when the pool of sign instances in the lexicon is expanded or different sets of body joints are considered.

5. Can a tool be developed to accurately measure and visualize variations in dominant hand trajectories between different sign languages, and how effective is this tool in identifying true and false friends across languages?

1.5 Main Contributions and Deliverables

In this section, we describe the main contributions and deliverables of this study. The contributions will be organized by chapters that follow this introduction. Each chapter corresponds to one article. Two of the chapters have been published in the proceedings of sign language conferences, one in a digital humanities journal and the last in a sign language studies journal.

In this thesis, we address machine and deep learning methods, tools and applications for sign language processing, recognition and comparison. The multi-modal nature of sign languages imposes specific analytical solutions to address this type of data. Furthermore, this thesis narrows the focus to methods designed for sign language corpora and lexica. Whether someone is comparing sign languages or develops applications to translate a sign language to a spoken language or vice versa, there are numerous aspects that can be measured and analyzed.

Domain experts gather information and compile them into data collections while data and their analysis can further inform and enlighten the expert [116, 65]. In our case, domain experts are sign language researchers and linguists. Greater information is offered, and these domain expert gains more in-depth understanding of the subject in matter. This loop results in greater insights and decisions [65].

This symbiotic relationship is at the core of the design process of the methods and tools presented in this thesis. The tools, for example, developed for supporting sign language annotation have been developed in such a way to allow faster processing of sign language collections and corpora, specifically those at an early compilation stage. That way, sign language researchers can use our pipeline to first recognize signing and gestural sequences which they can later on annotate at a specific gestural and sign level (instead of sequences of them). That data can then be used to train other models that can recognize phonological features such as handshape, location and orientation, to name but a few. Thereby, we argue that only by enhancing this cooperative relationship between sign language researchers and data it will be possible to compile large-scale sign language data-sets.

Our research resulted into the following tools and deliverables:

Manual Activation and Handedness Classifier

The purpose of this project is to provide a collection of tools to facilitate the annotation of signs and their formal features. The first tool allows the prediction and automatic annotation of sign and gestural sequences from video material as well as a general pipeline to allow researchers to re-train the model to fit a specific video and sign language material. The second application determines if a sign or gesture is one-handed or two-handed while the final tool attempts to detect the various handshapes as seen in a video [62].

Find the sign tool

The goal of this project is to investigate a reverse search method that allows a user to sign a query sign in front of a webcam and obtain a list of matching signs. In addition to current techniques based on (spoken language) glosses or phonological features, such as handshape and location, it offers a new approach of searching sign language dictionaries which requires no training to a particular sign language. Furthermore, it investigates whether different degrees of sign language proficiency can influence the results of the recognition process [61].

Sign and Search tool

This tool is an extension of the "find the sign" tool with respect to the experimental setup and total number of signs included in the lexicon. Different joint configurations, as detected by OpenPose [27], as well as four machine and deep learning techniques have been investigated as metrics to assess their effectiveness towards an effective sign suggestion ranking system [63].

DistSign tool

This work presents the DistSign tool which can be used to measure and visualize the variation in the trajectories of the dominant hands' wrist between the lexica of two sign languages. Using two sign languages which are assumed to be historically related, namely American Sign Language (ASL) and Ghanaian Sign Language (GSL), we evaluate the efficacy of the tool to identify true-friends across languages [57].

1.6 Structure of the dissertation

Following this introductory chapter, this thesis presents a series of papers. These are papers that have been published and peer reviewed. The papers are presented in the form of self-contained chapters. Although being self contained chapters, one could divide the work into chapters on tools targeted at sign language corpora and those developed for sign language dictionaries.

Chapter 2, "Towards a User-Friendly Tool for Automated Sign Annotation", presents a set of tools developed to assist the annotation process of signs and their formal features. This paper was published at the Digital Humanities Quarterly journal [62].

Chapter 3, "Signing as Input for a Dictionary Query: Matching Signs Based on Joint Positions of the Dominant Hand", presents a new method to search sign language lexica using the path of the dominant hand wrist. An initial version of this research was presented at the DH Benelux 2020 conference [58]. The completed study was later published at the 9th Workshop on the Representation and Processing of Sign Languages [61].

Chapter 4, " Sign and Search: Sign Search Functionality for Sign Language Lexica" serves as a deeper exploration of the sign language dictionary search functionality introduced in Chapter 3. It expands on the initial framework by delving into the potential use of different combinations of body joints in sign language recognition. This chapter goes a step further to explore how four different machine and deep learning techniques can enhance this process.This paper has been submitted and presented at the 1st International Workshop on Automatic Translation for Signed and Spoken Languages [63].

Chapter 5, "Assessing an automated tool to quantify variation in movement and location: a case study of American Sign Language and Ghanaian Sign Language", introduces a web-based tool which allows users to measure and visualize variation in movement and wrist location between two sign languages. An initial version of this research was presented at the African Sign Languages Workshop during the 10th World Congress of African Linguistics (WOCAL 10) [59]. The full findings of this study have been published in the Sign Language Studies journal [57].

Finally, in Chapter 6, we provide a summary of the results and contributions proposed in this thesis, the limitations these contributions

hold as well as future study recommendations.