Signing Avatars - Multimodal Challenges for Text-to-sign Generation

Sylvie Gibet, Pierre-François Marteau

Signing Avatars - Multimodal Challenges for Text-to-sign Generation

Sylvie Gibet[1] and Pierre-François Marteau[1]
[1] IRISA lab., University Bretagne Sud, France

Abstract— **This paper is a positional paper that surveys existing technologies for animating signing avatars from written language. The main grammatical mechanisms of sign languages are described, and in particular the sign inflecting mechanisms in light of the processes of spatialization and iconicity that characterize these visual-gestural languages. The challenges faced by sign language generation systems using signing avatars are then outlined, as well as unresolved issues in building text-to-sign generation systems.**

I. Introduction

Sign languages (SL) are fully-fledge languages that characterize the identity and culture of the Deaf. They belong to the family of languages known as visuo-gestural for which information is emitted by gestures and perceived by the visual system. Thus, the deaf people, with the practice of this language, develop a dexterity in their gestures and in their visual perception, an acuity of representation of the space and an expressivity which is conveyed by the whole body and manual movements, as well as the facial expressions, mouthing and gaze direction. This is why these languages are by essence multimodal. In practice, all body segments participate in spreading the message. The hands are the main vector of information transmission, but it is necessary to add to them certain gestures and body movements as well as facial expressions which are essential to qualify affectively an utterance, or to express negation or questioning.

SL challenge the usual boundaries of linguistic theories associated with oral languages. Like any language, they have a capacity for expression and abstraction which is based on their own linguistic structure with their vocabulary, grammar and semantics. But, unlike other oral languages, they integrate powerful mechanisms of spatialization and iconicity in their processes of sign formation. Furthermore, signed messages are generated and conveyed by movements and are perceived and interpreted by the interlocutor who recognizes the coded components of the gestural language, combined in space and time. This requires that the gestures be well articulated and visually understandable. In addition, dialogue between signers is made possible because an utterance is signed at a rate similar to that of spoken languages. Thus, spoken or written text can be translated into a continuous stream of movements that are perceived as natural and understandable by deaf interlocutors.

However, animating virtual signers has revealed to be a tedious task, mostly for two reasons: *i)* Although linguistic work on SL has led to a better understanding of the grammatical mechanisms of signed languages, computational linguistics in sign languages only deals with specific aspects of SL. *ii)* Animation methods are challenged by the complex nature of gestures involved in signed communication. This paper focuses on these two classes of inseparable problems, with more emphasis on the former, the latter having been addressed in [11].

In fact, the mechanisms underlying the generation of sign language utterances from written texts highlight two levels of generation. At the language generation level, there are a limited set of grammatical mechanisms that provide an economy of representation, based on the elementary constituents of signs (also called the phonological components) and grammatical inflectional rules [24]. These rules mainly rely on spatial and iconic representation of gestures that are present at all linguistic levels, from the phonological level to the lexical, syntactic or semantic level. Concerning the control and animation of the virtual avatar, it is mandatory to consider that signs differ sensibly from other non-linguistic gestures, as they are by essence multichannel. As mentioned above, each channel of a single sign conveys meaningful information corresponding to identified values of the phonological components. Therefore, combining and varying spatially and temporally these components expressed in parallel can serve as grammatical modifiers. Then, the combination in space and time of two or more components or signs is also possible to express concisely ideas or concepts. This intricate nature is difficult to handle with classical animation methods, that most of the time focus on particular types of motions (walking, beating, etc.), and that do not exhibit a comparable variability and subtleties. Therefore, the generated animated sentences have to be coherent from a motor coordination point of view, while preserving the linguistic coherence of the signed sentences.

This paper is a positional paper that reviews existing technologies for animating signing avatars from written language. From a descriptive grammar of SL perspective, we present the main challenges and unresolved issues for building complete translation systems from written or spoken language to SL. After describing the main existing signing avatar systems, we present the mechanisms of SL generation and grammatical inflections, relying on specific spatialization and iconicity dynamics. Then, for systems translating written text into signs, a set of unresolved problems are exposed, from computational languages to virtual character animation problems, including the building of sufficiently large annotated datasets.

II. Related Work

We first review some works about sign language text-to-sign translation systems and signing avatars technologies, and the related linguistic foundations. Please refer to [26]

for an exhaustive review on the subject. Fig. 1 shows in chronological order some existing signing avatars.

A. Linguistic representations

The first works on the phonology of the sign gave rise to different types of representations. Among these, the work of Stokoe [29] led to the description of ASL (American Sign Language) as a combination of elementary units constituting the signs: the location of the sign (*hand placement*), the shape of the hand (*hand configuration*) and its movement (*hand movement*). One of the basic assumptions is that two distinct signs can be differentiated when only one of the constituent parameters is changed (minimal pairs). A dictionary of ASL has been built from this representation. Continuing Stokoe's work, other parameters that participate in sign formation and distinction have been identified and added: *hand orientation* and *non-manual features*, including facial expression and gaze direction. According to these linguists, the phonological elements are arranged and synchronized spatially and temporally to form signs and utterances in SL.

Later, the *HamNoSys* (*Hamburg Notation System* [28]) proposed a notation system that takes the previous parameters and transcribes the signs in a linear way using Unicode computer symbols. With a linguistic approach to ASL, Liddell & Johnson have defined a phonetic system based on the Posture-Retention-Transition-Shift model (*PDTS*) distinguishing on each channel – hand configuration HC, orientation FA, placement PL, non-manual features NM – static elements and dynamic transitional elements.

More recently, in computational linguistics, SL gestures have been described using formalisms ranging from scripts to dedicated gesture languages. The language *SiGML* [7] based on *HamNoSys* was developed to generate the animations of 3D avatars. This language was then extended by incorporating Johnson's PDTS model. The modeling language *Azee* is based on a geometrical formalization, it aims to represent syntactic statements in SL in a non-linear way [8].

These scripting languages allow to describe signs or statements in a very analytical and precise way. However, the specification of new signs can be very tedious. It can be noted that most of the specification languages integrate explicit temporal elements within their formalism, this is the case in *SIGML*, *EMBRscript* [15], or in *Azee* in which the key postures of the avatar are specified at pre-determined moments. On the other hand, the *QualGest* [20] language is based on an implicit time formalization, the synchronization between the movements of the different gestural modalities (arms, hands) being managed at the level of the animation controllers. A formalism, called *Partition/Constitute (P/C) model* [16], proposes a linguistic synchronization scheme which relies on a 2D representation of a 3D syntax tree. This model facilitates the visualization of both the temporal and channel axis, while dealing with the coordination of the different channels.

Several linguistic representations have been interested in the mechanisms of grammatical inflections of signs. In particular, the *ATLAS* [22] project in Italian Sign Language has integrated modifier processes that influence the parameters of the signs (placement, configuration, movements), as well as the size specifiers. Other approaches have also focused on inflected signs. This is the case of *HTML* for Spanish Sign Language, or *AZee-Paula* [8] for proform generation in ASL.

B. Animation systems for signing avatars

Among the technologies available to animate signing avatars, three main approaches can be found: the first one consists in animating the avatar from very few data (for instance a set of key frames, key targets or trajectories). In this case it is necessary to control the whole production pipeline in the smallest details, from the specification of the basic linguistic elements, their arrangement by means of a procedural or rule-based logic, to the synthesis of a sequence of skeleton poses. The second approach reuses the movement as a basic resource. The third approach combines the first two approaches, or uses a learning-based model to generalize from the data.

With animation systems that link a scripting language to an animation engine based on pure synthesis, it is possible to achieve precision goals in motion control. However, these systems perform robotic movements. Moreover, they allow the creation of a limited number of signs. Indeed, building a vocabulary of signs and statements in SL using such methods can be a time-consuming task in terms of specification. Finally, time management remains complex to implement, both at the level of signs (management of synchronization between sign components) and transitions between signs (management of coarticulation).

An alternative to these synthesis systems is to develop data-driven animation methods. In this case, the movements of a signer are captured by motion capture techniques (MoCap) that simultaneously record hand movements, body movements and facial expressions. For example, the *SignCom* [11] and *Sign3D* [12] systems have been used to animate signing avatars in French Sign Language (LSF) with natural and realistic movements based on real signers' movements. In the context of these systems, two databases were built: i) a raw motion database in which captured motions are stored as skeletal postures characterized by transformations applied to the joints, and ii) a semantic database that maps multi-level annotations of signs to motions. The animation system is based on a multi-track concatenative synthesis principle, each track being associated to a set of dedicated controllers (facial animation, gaze direction, body or hand animation). This system led to the editing and construction of new LSF utterances by: i) replacing signs or groups of signs; ii) instantiating stereotyped syntactic patterns; iii) or replacing phonological components (hand configurations, manual, torso and head movements, and facial expressions) [10], [25].

Data-driven methods facilitate the production of smooth and believable animations of SL avatars. They result in the replay of relatively long sequences of SL but also in the modification of pre-recorded sentences to produce new utterances. However, the manipulation and adaptation of movements to new contexts requires the consideration of

Fig. 1. Some signing avatars in chronological order a) Dicta-Sign [6] b) eSign [18] c) New-York City avatar [17] d) Paula [23] e) Rosetta [4]

elaborate linguistic processes in order to keep the coherence of the produced SL content, both at the level of animations and the intelligibility of SL sentences.

The two types of control (data-based and hand-crafted synthesis) can be combined, to become a so-called hybrid synthesis. It is indeed possible to replace synthesis from scratch by motion data previously recorded and annotated, or combine procedural methods with data through machine learning. This hybrid control gives some flexibility and variational possibilities in the generated signed sequences. These approaches have been developed in several research teams: coupling data-driven and machine learning for directional verbs study [17], coupling hand-crafted and kinematics methods *SLPA/Azee* [23], [8], or coupling data-driven and inverse kinematics for spatial inflecting signs [25].

C. From Text to SL using Neural Machine Translation

With the advent in deep learning, recent approaches in Neural Machine Translation (NMT) have been developed with great success. Most of the research work related to NMT systems focuses on producing text from video. To support this work, video-based corpora have been created, mainly in recording studios with one or more cameras. An overview of European corpora is presented in [19]. The RWTH-PHOENIX-corpus, used in many studies of video sign recognition, includes 1980 German Sign Language (GSL) sentences describing weather forecasts [9]. Recently, the ROSETTA project has led to the creation of the Rosetta-LSF [21] corpus which contains 3 hours of LSF in a journalistic domain.

Very few NMT approaches have been proposed to produce automatically SL from text. Among others, Bahdanau et al. [1] developed an effective attention mechanism to translate English to ASL applied on weather reports. The system proposed by Stoll et al. [30] generates continuous GSL video from spoken sentences in two stages. First, it translates text into pose using an encoder-decoder architecture with attention that solves a motion graph. Then a pose to video network combines a convolutional image encoder and a Generative Adversarial Network (GAN). This system constitutes a proof of concept that demonstrates the capability of the NMT to produce GSL video from text, using a minimal amount of data annotation and text for training, the skeletal pose being extracted from video automatically using OpenPose [2]. Although this system does not reach the performances of avatar based systems, in terms of resolution and expressiveness, the approach is already promising.

III. Grammatical Mechanisms for the Generation of Signs and Utterances

Unlike oral languages which use the auditory-oral channel, sign languages use gestural and visual information. This specificity is at the origin of the omnipresence of iconic and spatial mechanisms in sign languages. Iconicity was initially defined by Cuxac as the process by which the signer makes the lived experience iconic, metaphorical or imaginary [3]. This is characterized by the link of more or less close resemblance between the entities and actions of the real or imaginary world, referent locations, and the sign and utterances that relate to them. Another specificity of SL concerns the difficulty of separating the different linguistic levels – phonological, morphological, lexical, syntactic, and semantic – that are specific to oral languages [24]. Thus, in SL, any modification operated at the level of the constituent components of the signs (phonological components, by analogy with oral languages) can potentially alter the sense of a sign or that of the utterance itself.

In the following we present the main grammatical mechanisms related to the generation of signs or signed utterances. First, we recall the principles of the sign language phonology. Second, we explore some processes of grammatical inflection in LSF, which can be characterized by the modification of one or more phonological components, and result in the modification of the meaning of the sign or utterance.

A. Phonological Components of Signs

SL phonology gives a structure to SL, hence creating a bridge between oral and signed languages. In oral languages, *phonemes* are units of sound that compose words and enable to distinguish one word from another. In SL, the parametric approach states that a sign is composed of parallel discrete values taken by SL phonological components [29]. Millet in her grammatical description of LSF [24] only considers the three main parameters: *Hand Placement*, *Hand Configuration* and *Hand Movement*. Hand placement is the location of the hand in the signing space (i.e., the space around the signer in which the signer's discourse takes place) or on the body of the signer. The hand configuration is the global handshape. Hand movement represents the trajectory of the hand over time. Phonological parameters can be described with a certain level of abstraction that relies on a discretization

of the space (definition of semantized areas in the signing space), a finite set of manual configurations (41 for Millet), and of typical hand motion trajectories. The generation of "standard" signs requires the combination of the phonological components described above. These signs, deprived of any syntactic context, are generally always executed in the same neutral zone of the sign space. Nevertheless, they must be executed with great precision, as the modification of one phonological component generates a different meaning, as for example the sign [NATURAL] which becomes the sign [NOT NEEDED] by modifying the manual configuration, the placement and the movement remaining unchanged.

B. Sign Inflections for Sign and Utterance Synthesis

When synthesizing signs or utterances, the form of some signs will vary to take the context into account. This is called the *sign inflection*. Two types of inflections are distinguished: inflections due to the illustrative nature of SL often referred to as iconic dynamics and inflections using spatial referencing. Based on the descriptive grammatical theory of Millet's [24], we will describe below some of these mechanisms.

1) Spatial referencing: The placement of entities inside a scene, their referencing, or the creation of interactions between those entities can be achieved through the variation of the hand placement or of the motion trajectory. Three spatial referencing inflections are presented below.

Placement. The signing space is characterized by a set of symbolic areas that can be *semantized* (for example, specific discrete areas are defined for pronouns). The signs performed within an utterance use specific locations in this space, which may correspond to a specific linguistic function. At the lexical level, the process of *Spatialization* consists in placing a sign at a given location that is not that of its neutral anchor. At the syntactic level, the *Locus* represents a 3D location in the signature space that allows to refer to predefined entities in a discourse or to give them a relative placement with respect to others.

Pointing. Pointing can take different forms, either by carrying a meaning of its own (value of a pronoun for example), or by pointing to an entity. In the latter case, the pointed entity constitutes the meaningful information, it indicates the subject(s) or object(s) of an action, or it associates virtual objects with $3D$ locations in the signing space for future referencing of these objects. Designating a place (or locus) is often done by pointing to the index (or other manual configuration). It is also possible to do this by gaze orientation or shoulder gesture.

Indicating verbs correspond to signs that are inflected according to the agent/recipient relationship. The trajectory of the hand thus changes according to the agent (subject of the verb) and the recipient of the action. For example, the verb [GIVE] can be inflected according to different trajectories whose starting and ending points determine the pronouns representing the agent or the beneficiary. This is illustrated in Fig. 2. In the sentence "I'll give you a drink": the hand trajectory is a line that starts near the body (pronoun [I]) and ends in front of the signer (pronoun [YOU]). When the starting and ending target points are reversed, the sentence becomes "You give me a glass". The other pronouns are located in other areas of the signing space (locus to the right of the signer for the pronoun "He/She"), ellipsis/arc in front of the signer for the pronouns "we/you" and ellipsis to the right of the signer for the pronoun "they".

Fig. 2. In the sentence "I give you a glass", the hand movement follows a line, starting from an area near the body (pronoun [I]) and ending in front of the signer (pronoun [YOU]). In the sentence "You give me a glass", the starting and ending points are reversed

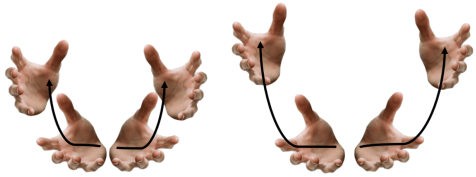

Fig. 3. Iconicity on the LSF sign *[BOL]* (bowl): the size of the bowl corresponds to the amplitude of the motion (image extracted from [26])

2) Iconicity: SL iconicity refers to the similarity between the sign and what it designates. Three illustrative mechanisms of iconicity are described below.

Size and Shape Specifiers are processes that allow to describe the shape or size of signed objects. They can be lexical signs, as for example the signs [BOL] and [VERRE] in LSF which can be inflected by adding an adjective to them, at the level of the shape (e.g., [VERRE-A-EAU], [VERRE-DE-CHAMPAGNE]), or at the level of the size (e.g., [GRAND-BOL], [PETIT-BOL] as shown in Fig. 3).

Static proforms represent animated entities (person, object) and are characterized by a limited number of configurations. They avoid naming an entity multiple times and make referencing these entities in space more efficient. For example, the [PERSON] proform can be quickly positioned in the narrative scene. Moreover, the person can be represented in different positions, which leads to different manual configurations (e.g., a raised finger for a standing person or a curved one for a sitting person). Also, several people can easily be represented in a space (around a table for example) or in a conference room.

Direct objects in indicating verbs. Some transitive indicating verbs can be inflected according to the direct object. In this case, the handshape is modified. For example, the sentences "I give you a glass" or "I give you a sheet of paper" are performed in LSF in the same way, except for the hand configuration that changes according to the direct

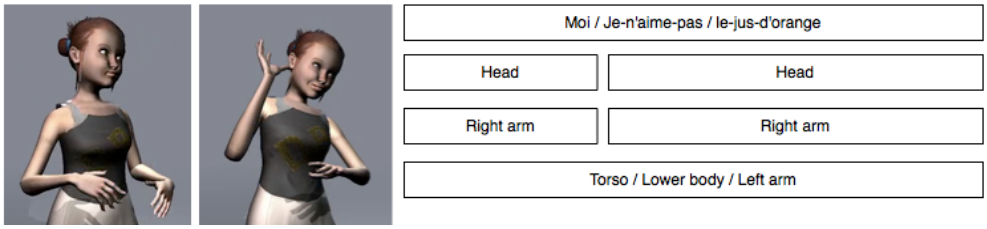

Fig. 4. Combining three signs: [PRO-1] / [LIKE-NOT] / [ORANGE-JUICE].

object [GLASS] (handshape "C", see Fig. 2) or [SHEET-OF-PAPER] (handshape "pinch").

3) Hand movements: Hand movements in SL can be characterized by linguistic features. We describe three of them below.

Finite set of motion paths. Typical motion paths are used in SL. Among them, there are simple elementary movements (Pointing, Line, Arc, Ellipse), or complex movements (Spiral, Waves, Lemniscate, etc.). These movements can be achieved in various locations (Locus), expressed in the signing space (starting and target points). They can be unitary movements, or repeated movements.

Dynamic Proforms. Some behaviors or gaits can also be represented by dynamic *proforms*. This is the case for example when we want to describe the gait of an animal (a chicken, a bear or a lion for example). The shape of the hand is modified to represent the shape of the animal's leg and its gait, which is more or less heavy. It is also possible to indicate by the movement of the hands the quality of the movement (flexibility, lightness, style).

Motion dynamics. The temporal and physical properties of the movements can also change the sense of the signs. For exemple, the signs [CHAIR] and [SIT DOWN] in LSF have the same manual configurations and the same spatial trajectories, but have different dynamics. It should also be noted that the way in which the contacts are executed (touch gently or hit) may change the meaning of the signs.

4) Facial expressions: Facial expressions are of primary importance in SL. They give not only information related to the quality of what is being expressed (emotion, prosody), but they also provide objective information that contributes to the semantics of the sentence.

Adjectives or adverbs. In SL, facial mimics may serve as adjectives (e.g., inflated cheeks make an object large or cumbersome, while squinted eyes make it thin). They can also be used as adverbs (e.g., blowing hard, mimicking the force of the wind), or indicate whether the sentence is a question (raised eyebrows) or a negation (frowning). It is therefore very important to preserve this information during facial animation.

Affects. Other facial expressions express affects that concern all or part of the sentence. A misinterpretation of these deliberately shown emotions can alter the meaning of the sentence. For example, if one reports an accident with a smile, it can be misinterpreted.

Clausal aspects. Facial expressions also give information about the clausal aspect of the sentence. A negation can be expressed by a manual sign or by a facial expression (frowns indicating negation at the end of the sentence), or both. Similarly, the interrogative sentence can be distinguished from the declarative sentence by the interrogative facial expression. In some conditional sentences it is also possible to use specific facial expressions such as the [PI] labialization or the eyebrow lift (e.g., "If it rains, I'll stay home").

5) Other non manual signs: Other linguistic features concerning the movements of the upper torso, the orientation of the head and the direction of gaze, also carry semantic information.

Future tense. Movement of the torso backwards or forwards makes it possible to place events in the past or the future. For instance, the trunk slightly bent forward can indicate an action performed in the future.

Change of role. The orientation of the torso can be used for the change of role. For example, in a story, if several characters are involved, they can be placed in the signing space. Changing from one character to another can be achieved by turning the torso alternately towards each character.

Eye gaze. Eye gaze is a modality used in SL to reference a specific object (referred entity in the signing space). It can also be used to improve the understanding of the sign, as in the sign [READ] which corresponds to the action of reading, and for which the eyes follow the movement of the fingers, as in the action of reading.

6) Combining the various components in the discourse: All of the grammatical mechanisms described above are used in the construction of discursive utterances (stories, dialogues, etc.) in SL. Standard signs and illustrative signs are arranged in the utterances, either sequentially or by modifying information on one or more phonological channels. In the example given in Fig. 4, the sentence "I like fruit juice" is transformed into the sentence "I don't like orange juice". The movement of the chest as well as the lower body and the left arm are preserved. However, the movements of the head and the right arm, as well as the facial expression are modified.

It should be noted that, depending on the sign, the hand movements can be performed with one hand only (the dominant hand) or with both hands. The hand movements can be totally symmetrical with respect to one of the three

planes – sagittal, longitudinal, horizontal –, or alternated with respect to one of these planes, or they can be realized in an asymmetrical way with a dominant hand that establishes the basis of the sign and a dominant hand that performs the main action, as in the example "The plane takes off", where the static left hand (flat form), represents the runway and the dominant right hand the upward movement of the plane.

IV. Challenges for Sign Language Translation and Avatar Technology (SLTAT)

Since a few years, the SLTAT community has mastered the issue of synthesizing isolated signs. The main purpose of this synthesis is to build signs for bilingual dictionaries or educational tools for learning SL lexicon. In this case, the signs are not contextualized and have to be executed in their standard form (with stable phonological components). The synthesis of utterances is mainly used by machine translation systems to accurately express a spoken or written sentence into a given SL and for storytelling [5]. Whether for translating or storytelling, the signs that compose the sentences are influenced by the context, and both standard and inflected signs must be used within the sequence of signs. Indeed, the simple concatenation of standard signs is not sufficient to take into account all the grammatical mechanisms mentioned above, and to produce correct statements in SL.

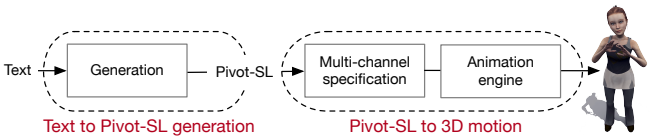

Fig. 5. FromText to Pivot-SL generation, and from Pivot-SL to 3D motion

Fig. 5 schematizes the challenges still encountered for signing avatars. It illustrates the two main blocks required to animate a signing avatar: 1) the *Text to Pivot-SL generation* block involves the translation of the written text in a natural language – possibly generated from audio signal – into the computational linguistic representation of SL (so-called *Pivot-SL*); 2) The *Pivot-SL to 3D motion* block represents the translation of the symbolic SL representation into the multichannel specification that produces a continuous 3D motion stream, thanks to the animation engine.

Although much work has been done in this area of research, there are still many unresolved challenges in SL utterance synthesis, which concern both the formal linguistic representation of SL and the animation system adapted to this representation. We describe below these challenges related to the different steps.

A. Formal linguistic representations of SL

In order to account for the grammatical processes underlying the formation of signs and utterances, it is necessary to introduce at the formal level a level of parameterization that incorporates the various inflectional mechanisms presented above, as well as the logics of spatio-temporal representation, and the coordination/synchronization rules.

Inflection representation. A linguistic formalism must incorporate the various inflectional processes that cover the iconic aspects of SL. For instance, indicating verbs, whether transitive or not, take different parameters within a sentence, representing the agent (subject of the action), the patient (beneficiary of the action), and possibly the direct object. In the sentence "I give him a book", the subject is the first pronoun, the beneficiary the 3rd one, and the direct object is a book, represented by a "C" hand configuration.

Another example concerns proforms which are intuitive and powerful syntactic tools, particularly adapted to play the role of pronouns in sentences, to reference lexical items, or to describe situations of relative spatial positioning. For example, a pencil can be represented by the index finger and it is possible to use this manual configuration to position it in a pot or on a table. Proforms are an efficient way to describe a wide variety of situations while avoiding explicit reference to the entities they represent.

Spatial representation. The signing space is the place where the signer places the entities of his discourse. Within this space physically limited by its extensional capabilities, the signer can describe an infinite and constantly changing space. Thus, the signing space requires the identification of abstract and discretized areas that can then be referenced in the utterances. These spatial references, absolute or relative, must be carefully defined. Other spatial mechanisms, should also be considered [24]. Among them, we can cite the multiple arrangement of objects, which involves both horizontal alignments (*sweeping*) and vertical stacks (*stacking*). The relative placement is another frequent mechanism in LSF, allowing objects to be iconically referenced to each other (e.g., object 1 is on (in) object 2). It can be achieved by static proforms, thus ensuring the syntactic consistency of the sentence [24].

Coordination/synchronization rules A common way to analyze a sentence structure in natural language processing for oral languages is to display it in the form of a syntax tree or graph. However, the multichannel aspect of SL makes it difficult to describe an SL utterance as a graph. A formalism has therefore to be designed to represent both the sequence of signs (manual and non manual signs) and the body channels. Such a representation makes it possible to manage the coordination and the synchronization of the various channels.

Timing rules Time management for multichannel control of avatars can be achieved in different ways. It can be (i) specified explicitly, in a relative or absolute manner, or (ii) implicitly, just indicating the sequence of motion chunks on the different channels. (iii) A third possibility could be to allow the system to respond reactively to external events, or to anticipate the movement of certain body parts during complex tasks. In the latter case, the language could be based on the representation of spatio-temporal events that would mark anchor points at the inter and intra-channel levels.

B. Animation systems for signing avatars

Multichannel coordination is one of the main challenges

Fig. 6. Hand animation with 3D mesh constraints (image extracted from [27])

of animation systems for SL. It involves the spatial ordering and the temporal synchronization of the movements generated on their respective channels (either by extraction from the database or by synthesis) in order to respect the spatio-temporal *patterns* of the signs. It is likely that the different motion elements have not the same duration. The consequent problem is twofold: *i)* a common timeline has to be found, possibly as the result of a combinatorial optimization, or driven by linguistic rules. Up to our knowledge though, no existing model of SL describe such temporal rules or model the synchronization of the different channels *ii)* once a correct timeline has been devised, the temporal length of the motion chunks has to be adapted, while preserving the dynamic of the motions. To this end, time warping techniques can be used [13]. However, inter channels synchronizations may exist (for example between the hand and the arm motions [14]).

Coarticulation effects are characterized in SL by the influence of one sign on the adjacent signs. It can be expressed both in the transitions between the signs and within the sign, to take into account the previous and following signs. When animating signing avatars, it is essential to manage coarticulation, and this has to be achieved at the different levels (intra and inter channel) in order to take into account the contextual information of the signs in the generated sequence. Incorporating spatio-temporal variability in the motion signals can be used to enhance the overall expressiveness and style of the virtual signer. However, small spatial or temporal variations can profoundly alter the meaning of a sentence. In the near future, these spatio-temporal patterns will be retrieved from the data, through machine learning approaches.

Morphological adaptation. Most signing avatars use signed data of a single signer. The application to other avatars involves processes of *retargeting* that can transfer the movements to other characters with different morphologies (man, woman, child, even animal). Moreover, the constraints of the SL are linked to the content of the signs and statements. Indeed, many signs involve contact between the two hands, or between each hand and a part of the body. All these constraints must be precisely specified in the signer's space (e.g. the hands must remain above the table) or expressed in a qualitative way (e.g. "the index finger must touch the shoulder"). Finally, hand animation is particularly important in SL, as it requires high precision and avoidance of inter-penetrations, hence the consideration of spatial constraints in the optimization processes [27] (see Fig. 6).

SL Corpus. Data is at the heart of the technologies and methods used for signing avatars. Three categories of data are available: video, motion capture and annotations. Several issues arise in defining the corpus. The first one concerns the trade-off between breadth and depth of the corpus. If the objective is to have a lexicon that covers a large domain, including several themes, an extensive corpus will be preferred. If, on the contrary, the objective is to have a limited vocabulary and to reuse it in different sentences, then the in-depth approach will be chosen. In this case, many instances of the same signs with variations will be considered in the predefined vocabulary. The second issue concerns the nature of the variations themselves that should be included in the corpus for editing, synthesis and recognition. Several levels of sign inflections can be considered, including: i) contextual variations, for example by varying the predecessors and successors of the same sign, thus facilitating the study of coarticulation; ii) modulations of signs at the level of their phonological components, by changing for example placements, hand configurations or hand movements ; iii) spatial and iconic variations that make possible the specific study of inflectional mechanisms such as size-and shape specifiers, pointing, modulations of indicating verbs according to pronouns and direct object; iv) variations in style or prosody which induce kinematic modifications of the movements (more or less rapid, fluid or jerky, etc.). Finally, an essential concern in the construction of the corpus is the acted or spontaneous quality of the movements produced by the signing actors.

Perceptual evaluation of signing avatars. The quality of the motion generated by text-to-SL systems needs to be evaluated perceptually. Different evaluations have already been performed [17], [11], both animation and SL being intertwined. First, the different modalities can be tested separately – facial expressions, gaze direction, hand motion –, according to the meaning sought. Then, the ability of the generating system to produce a realistic and comprehensible sign or utterance can be evaluated perceptually, on the one hand with criteria of precision and naturalness, and on the other hand of intelligibility and grammatical understanding.

C. Neural Machine Translation for text to sign generation

As suggested by early approaches mentioned in the state of the art, systems for automatic translation from one spoken language to another open the possibility of automatically translating a spoken language to a SL, using deep learning. However, although NMT-based methods have been very successful in natural language translation tasks, machine translation systems for translating from text to a formal or visual representation of SL are still at an embryonic stage. To our knowledge, [30] is one of the rare attempts in this area. Indeed, in addition to the technical difficulties related to the inherently multi-channel nature of SL, sequence-to-sequence approaches that have proven very effective for oral language translation struggle to solve the problem for SL, for several reasons. The MoCap data available for training neural network models is limited. However, there are many small

corpora, primarily video corpora, that can be merged and adapted across domains to provide large training datasets. Nevertheless, these corpora may be insufficient to cover the large spatial variability due to the many grammatical inflections of SL sentences. Moreover, compared to oral languages that encode linearly both the written word – as a sequence of phonetic elements – and their corresponding sound units, there is a lack of aligned resources between text and SL required to provide parallel resources for training models. To overcome this lack of labeled motion data, the solution could be to go through an intermediate language that would guide the learning process by introducing linguistic knowledge. However, the lack of a commonly accepted written representation for SL does not facilitate the recording of sufficient information to match a text in an oral language to its transcription in SL.

V. Conclusion

In this article, we have highlighted a non-exhaustive set of linguistic mechanisms specific to sign languages and described some technological challenges for the production of utterances by means of dedicated linguistic representations and animated signing avatars. If recent works already apprehend part of the SL linguistic modeling and have developed efficient animation systems for signing avatars, the huge variability that characterizes these sign languages and the complexity of the grammatical inflecting mechanisms, relying on the implementation of dedicated linguistic models and animation controllers, open up research avenues that are still little explored.

In the near future, the possibility of capturing large volumes of data and the advent of deep machine learning models will make it possible to design autonomous translation systems, from text to sign or sign to text, based directly on annotated video data. More broadly, this opens perspectives towards automatic translation systems from oral languages to sign languages, or vice versa.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[2] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:1302–1310, 12 2018.

[3] C. Cuxac. *La langue des signes française (LSF) : les voies de l'iconocité (French) [French Sign Language: the iconicity ways]*. Faits de langues. Ophrys, 2000.

[4] B. Dauriac, A. Braffort, and E. Bertin-Lemée. Example-based multilinear sign language generation from a hierarchical representation. In *7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 21–28, Marseille, France, June 2022. ELRA.

[5] S. Ebling. *Automatic Translation from German to Synthesized Swiss German Sign Language*. PhD thesis, University of Zurich, 2016.

[6] E. Efthimiou, S.-E. Fotinea, T. Hanke, J. Glauert, R. Bowden, A. Braffort, C. Collet, P. Maragos, and F. Lefebvre-Albaret. The dicta-sign wiki: Enabling web communication for the deaf. In K. Miesenberger, A. Karshmer, P. Penaz, and W. Zagler, editors, *Computers Helping People with Special Needs*, pages 205–212. Springer Berlin Heidelberg, 2012.

[7] R. Elliott, J. R. Glauert, J. Kennaway, I. Marshall, and E. Safar. Linguistic modelling and language-processing technologies for avatar-based sign language presentation. *Universal Access in the Information Society*, 6(4):375–391, 2008.

[8] M. Filhol, J. McDonald, and R. Wolfe. Synthesizing sign language by connecting linguistically structured descriptions to a multi-track animation system. In *International Conference on Universal Access in Human-Computer Interaction*, pages 27–40. Springer, 2017.

[9] J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney. Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus. In *Language Resources and Evaluation*, pages 3785–3789, Istanbul, Turkey, May 2012.

[10] S. Gibet. Building french sign language motion capture corpora for signing avatars. In *Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, LREC 2018*, Miyazaki, Japan, May 2018.

[11] S. Gibet, N. Courty, K. Duarte, and T. Le Naour. The signcom system for data-driven animation of interactive virtual signers : Methodology and evaluation. In *ACM Transactions on Interactive Intelligent Systems*, volume 1, 2011.

[12] S. Gibet, F. Lefebvre-Albaret, L. Hamon, R. Brun, and A. Turki. Interactive editing in french sign language dedicated to virtual signers: requirements and challenges. *Universal Access in the Information Society*, 15(4):525–539, 2016.

[13] A. Héloir, N. Courty, S. Gibet, and F. Multon. Temporal alignment of communicative gesture sequences. *Computer Animation and Virtual Worlds*, 17:347–357, July 2006.

[14] A. Héloir and S. Gibet. A qualitative and quantitative characterization of style in sign language gestures. In *Gesture in HCI and Simulation, LNAI 5085*, pages 122–133. Berlin:Springer, 2008.

[15] A. Heloir and M. Kipp. Real-time animation of interactive agents: Specification and realization. *Applied Artificial Intelligence*, 24(6):510–529, 2010.

[16] M. Huenerfauth. *Generating American Sign Language classifier predicates for English-to-ASL machine translation*. PhD thesis, University of Pennsylvania, 2006.

[17] M. Huenerfauth, P. Lu, and H. Kacorri. Synthesizing and evaluating animations of american sign language verbs modeled from motion-capture data. *6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 22–28, 2015.

[18] J. R. Kennaway, J. R. W. Glauert, and I. Zwitserlood. Providing signed content on the internet by synthesized animation. *ACM Trans. Comput.-Hum. Interact.*, 14(3):15, 2007.

[19] M. Kopf, M. Schulder, and T. Hanke. Overview of Datasets for the Sign Languages of Europe, July 2021.

[20] T. Lebourque, S. Gibet, and P. Marteau. High level specification and animation of communicative gestures. *Journal of Visual Languages and Computing*, 12:657–687, 2001.

[21] LIMSI and LISN. rosetta-lsf, 2022. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

[22] V. Lombardo, F. Nunnari, and R. Damiano. A virtual interpreter for the italian sign language. In *International Conference on Intelligent Virtual Agents*, pages 201–207. Springer, 2010.

[23] J. McDonald, R. Wolfe, J. Schnepp, J. Hochgesang, D. G. Jamrozik, M. Stumbo, L. Berke, M. Bialek, and F. Thomas. An automated technique for real-time production of lifelike animations of American Sign Language. *Universal Access in the Information Society*, 15(4):551–566, 2016.

[24] A. Millet. *Grammaire descriptive de la langue des signes française: dynamiques iconiques et linguistique générale*. UGA Editions, 2019.

[25] L. Naert and C. Larboulette. Motion synthesis and editing for the generation of new sign language content. *Mach. Transl.*, 35(3):405–430, 2021.

[26] L. Naert, C. Larboulette, and S. Gibet. A survey on the animation of signing avatars: From sign representation to utterance synthesis. *Comput. Graph.*, 92:76–98, 2020.

[27] T. L. Naour, N. Courty, and S. Gibet. Skeletal mesh animation driven by few positional constraints. *Computer Animation and Virtual Worlds*, 30(3-4), 2019.

[28] S. Prillwitz. *HamNoSys: Version 2.0; Hamburg Notation System for Sign Languages; An Introductory Guide*. Signum-Verlag, 1989.

[29] W. C. Stokoe. Sign language structure: An outline of the visual communication systems of the american deaf. *Studies in Linguistics, Occasional Papers*, 8, 1960.

[30] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden. Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, pages 1–18, 2020.