Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2488–2497
Marseille, 20-25 June 2022

A Low-Cost Motion Capture Corpus in French Sign Language for Interpreting Iconicity and Spatial Referencing Mechanisms

Clémence Mertz*, Vincent Barreaud*, Thibaut Le Naour†, Damien Lolive*, Sylvie Gibet‡
*Université Rennes1, IRISA, France clemence.mertz, vincent.barreaud, damien.lolive@irisa.fr
†Motion-Up, France contact@motion-up.com
‡Université Bretagne Sud, IRISA, France sylvie.gibet@irisa.fr

Abstract

The automatic translation of sign language videos into transcribed texts is rarely approached in its whole, as it implies to finely model the grammatical mechanisms that govern these languages. The presented work is a first step towards the interpretation of French sign language (LSF) by specifically targeting iconicity and spatial referencing. This paper describes the LSF-SHELVES corpus as well as the original technology that was designed and implemented to collect it. Our final goal is to use deep learning methods to circumvent the use of models in spatial referencing recognition. In order to obtain training material with sufficient variability, we designed a light-weight (and low-cost) capture protocol that enabled us to collect data from a large panel of LSF signers. This protocol involves the use of a portable device providing a 3D skeleton, and of a software developed specifically for this application to facilitate the post-processing of handshapes. The LSF-SHELVES includes simple and compound iconic and spatial dynamics, organized in 6 complexity levels, representing a total of 60 sequences signed by 15 LSF signers.

Keywords: Corpus, Motion Capture, French Sign Language, Kinect Azure, MotionUp software

1. Introduction

Assistance in learning French sign language (LSF) requires the creation of appropriate resources and the development of digital interfaces that facilitate access to information. However, main existing data are limited lexicons, a few narrations or dialogues in LSF, and very few are publicly available and labeled with transcribed texts. Furthermore, corpora dedicated to grammatical structures in LSF remain almost non-existent.

Our goal is to design and develop digital and pedagogical tools that facilitate the learning (or reinforcement) of LSF, focusing on well identified grammatical mechanisms. With the increase in the amount of data and the advent of new machine learning techniques, these tools should assist the automatized translation from videos in LSF into written transcriptions in French.

Since it is not possible to tackle the complete machine translation pipeline, given the successive transformations, from video to 3D motion, and from 3D motion to written LSF, we focus in this work on a partial modeling and interpretation system that makes possible the interpretation of signed sentences. This system is characterized by 1) a simplified methodology of the translation process; 2) the selection of limited LSF grammatical mechanisms; 3) the design of an original data acquisition technique that facilitates the capture, storage and post-processing of large volumes of data.

The machine translation process can be described as a dual translation process, in which the signed and labeled videos (the subtitles being usually in French) are transformed into precisely annotated 3D movements in the form of a sequence of 3D skeletal poses,and these 3D movements are then translated into written transcriptions (called glossed-LSF), using an intermediate “pivot” language that takes into account the spatial and structural modeling of the sign language. Then the glossed-LSF is translated into written French. We will focus here on the transformation from 3D motion to glossed-LSF.

Unlike oral languages, sign languages use gestural and visual information. This specificity is at the origin of the omnipresence of spatial and iconic mechanisms in sign languages. Iconicity is characterized by the more or less close resemblance between the sign and what it designates (Cuxac, 2000). These iconic structures are essential in situations of scene description or storytelling. In our corpus, we have chosen to include several grammatical mechanisms of iconicity and spatial referencing. The main objective is to give the largest representation of these mechanisms, with a corpus focusing exclusively on it, other issues such as lexical recognition being minimized. Moreover, we propose to implement each utterance of the corpus with a large number of lexical or grammatical inflections, i.e. by using various methods to specify a location, or by varying the shapes of signs or motion trajectories. In order to limit the study, we have not treated facial expressions which convey other forms of syntactic or semantic inflections.

The need to capture large volumes of motion data (from 3D captured motions) led us to a new motion acquisition technology and protocol characterized by: (i) a low-cost rather than high-resolution motion capture device, thus limiting hand data post-processing; (ii) the possibility to perform the recordings at the deaf person’s home and not in a motion capture studio; (iii) as hand processing remains a real challenge in such data sets, we relied on a specialized software to ”paste” hand configurations and manually edit the recorded motion data.

This paper describes the LSF-SHELVES corpus that implements the spatial referencing and iconicity mechanisms in LSF, selected through examples of positioning objects in relation to each other on shelves. It proposes to produce utterances with increasing levels of difficulty. The original low-cost motion acquisition technique used to construct the dataset and then supplement it with editing techniques, including incorporating manual configurations and correcting wrist orientation trajectories, is then detailed. Section 2 reviews existing video and motion capture data sets for sign language recognition systems. Section 3 describes the objectives and the content of the LSF-SHELVES corpus. Section 4 presents the data acquisition process. Section 5 details the post-processing. Finally, Section 6 concludes and gives perspectives to this work.

2. Related Work

Given the visual and gestural nature of sign languages, most corpora have been built using video cameras or motion capture technologies.

Being low-cost and easy to collect, video-oriented corpora usually constitute the prime material in linguistic analysis. For instance, (De Beuzeville et al., 2009; Ormel et al., 2017) have collected video data, focusing on directional verbs and the use of space. Studies on coarticulation have given rise to different corpora (Ormel et al., 2013; Ojala et al., 2009). In French Sign Language, it is possible to study iconicity and role shift with the LS-COLIN corpus (Cuxac et al., 2002), and classifier predicates in narratives or stories (Millet, 2006; Millet, 2019). A remarkable work of comparison of video corpus recorded since 2012 is available on the website of the University of Hamburg (DGS-Korpus, 2021).

In the case of sign language recognition, many works also use video recording. Early approaches focus on the separate recognition of the basic components of sign languages, namely handshapes and hand movements. Prior work on handshape recognition (Yuntao and Weng, 2000; Lu et al., 2003; Vogler and Metaxas, 2004; Isaacs and Foo, 2004; Ding and Martinez, 2009) used a finite set of handshapes linguistically identified in the specific sign language. More recently, handshape recognition applied on a large vocabulary of hand poses was achieved, using convolutional neural networks (Koller et al., 2016).

Research on sign language recognition has also explored hand motion trajectories, which was considered as important for sign recognition (Junwei et al., 2009; Dilsizian et al., 2016; Pu et al., 2016). Other research has tackled the problem of full sign recognition for isolated signs. However, the recognition rates for about 1000 signs across multiple signers did not give the expected results (accuracy around 70%).

More recently, new techniques based on deep learning and relying on data-driven end-to-end approaches have targeted sign recognition in a continuous stream of sign language (Cui et al., 2017). To support this work, new video-based corpora have been created, mainly in recording studios with one or multiple cameras. An overview of the European corpora is presented in (Kopf et al., 2021). The corpus RWTH-PHOENIX-Weather includes 1980 German sign language sentences describing weather forecasts (Forster et al., 2012). It is used in many studies on video sign recognition, mostly relying on neural networks, such as in (Konstantinidis et al., 2018; Forster et al., 2018; Huang et al., 2018; Pu et al., 2019; Cihan Camgöz et al., 2020).

In those previous work, recognition based on video input did not involve intermediate 2D or 3D poses. Recent work adds this step of extracting 2D or 3D vectors from frame-by-frame video data, (Cao et al., 2021), which provides a skeleton estimation. However, most of these promising approaches do not consider hand reconstruction, because of the high dimensionality of the hands and the numerous occlusions encountered, especially in sign languages. Studies using pose estimator and hand reconstruction for sign language recognition can be found in (Li et al., 2020b), relying on the (Li et al., 2020a) corpus (isolated ASL sign corpus including 2000 common signs), in (Metaxas et al., 2018) with the corpus (Athitsos et al., 2008) (isolated ASL sign corpus including 5000 isolated signs), in (Belissen et al., 2020) with the corpus Dicta-Sign-LSF-v2 (Efthimiou et al., 2010) with the hand model of (Koller et al., 2016), and in (Konstantinidis et al., 2018) with the corpus RWTH-PHOENIX-Weather.

Recognition from skeletal data can meet the requirements of sign languages that demand high precision in hand movements. Such an approach was achieved, either through hand movement tracking systems (usually two trackers, one for the body and one for the hands) (Vogler and Metaxas, 2004), using Ascension Technologies MotionStar™alongside Virtual Technologies Cyberglove™ (Cooper et al., 2011), referring to the Polhemus tracker (Waldron and Kim, 1995), or relying on MoCap technology. Several MoCap corpora in French sign language have been built such as (Benchiheub et al., 2016) without the hands or (Gibet, 2018; Naert et al., 2020) with hands included, but they are dedicated to the synthesis of utterances in sign language using 3D avatars, and they include very little variability on iconic and spatial mechanisms.

In the studies cited above, sign language recognition is primarily a recognition of signs, isolated or in data streams, but very few of these studies aim at recognizing grammatical sentences. This type of recognition relies essentially on the collection of dedicated data. Moreover, the required datasets need to be large enough to be suitable to deep learning solutions. The collection of such a large dataset, representative of the targeted grammatical mechanisms, with a high variability, requires on the one hand to minimize the post-processing that can be tedious, and on the other hand to have a large number of volunteers able to sign the sentences in LSF. Therefore, a light-weighted and convenient protocol is useful for this purpose. It is centered on a portable device, which allows the capture process to be moved to the deaf. This constitutes a trade-off between data accuracy – that video corpora cannot achieve – and the heaviness of the MoCap protocol. In this paper we present both our LSF-SHELVES corpus and a low-cost and innovative solution to acquire this corpus.

3. Corpus Definition

3.1. Motivation

As introduced in section 1, the goal of our corpus is to provide material for future work on iconicity and spatial referencing in LSF, knowing that these aspects are omnipresent in sign languages. More specifically, this corpus is intended to analyze, model and recognize the elements of LSF representing grammatical structures. The research objective is twofold. On the one hand, this corpus will constitute a basic resource for linguistic analysis. On the other hand, it will be used for the modeling of signed sentences, facilitating their interpretation and the recognition of written glossed-LSF, using recent machine learning approaches. Grammatical structures have been chosen on the following criteria: (i) It is highly challenging for an automated tool to recognize referencing structures and extract information from them. Indeed, the variability of the referencing mechanisms raises issues never addressed in sign-based recognition systems; (ii) Moreover, as shown in (Metaxas et al., 2018), we will show the importance of introducing linguistic knowledge into the automatic recognition process. Following those criteria, spatial referencing seems to be a consistent subset of grammatical mechanisms with the particular strength of being ruled by linguistic guidelines. Based on the descriptive grammatical theory of Millet's (Millet, 2019), we include in our corpus some descriptions with iconic mechanisms, which essentially operate at lexical and syntactic levels. According to Millet, these mechanisms directly influence the three phonological components that are Hand Placement, Hand Configuration and Hand Movement (Stokoe, 1960). These basic linguistic mechanisms are detailed below, in the Placement and Hand Shape sections. More complex sentences use these linguistic mechanisms to describe multiple objects positioning, relative referencing, and sophisticated geometrical shapes. Finally, our corpus incorporates dynamic descriptions involving moving hand configurations.

3.2. Linguistic Mechanisms

3.2.1. Placement

In LSF, lexical signs performed during an utterance have a specific location in the signing space, i.e. the space immediately surrounding the signer. There are several types of placements, each one corresponding to a specific linguistic function (Millet, 2019). At the lexical level, the mechanism of **Spatialization** consists in placing a sign in a given place that is not that of its neutral anchoring place. At a syntactic level, the **Locus** represents a 3D location in the signing space. With these 3D locations, it becomes possible to refer predefined entities in a discourse or to give them a relative placement with respect to others. The Locus can be activated by different means, for example by a pointing gesture or a pointing through gaze or upper-body motion. In our corpus, we will only consider index pointings. Three different situations are considered:

Sign and place. In this situation, the entity is signed in its neutral anchoring location in front of the signer and then is placed at a precise location. In Figure 1 a., for example, the bowl can be signed in its neutral lexical space and then placed on the left of the shelf, while keeping the hand shape to depict the bowl.

Sign at place. For lexical signs not anchored on the body, it is possible to go to the specific *Locus* and to directly sign the entity at this location. For example, in Figure 1 a., The bowl can be signed at the 3D *Locus*.

Point to entity. Pointed target is retained in this category of simple placement, although it implies both the Pointed Locus and the hand movement gesture. In our corpus, we will consider index pointing to a specific *Locus* in the signing space to indicate the exact position of the entity. For example, in Figure 1 a., a pointing gesture can be used to indicate the sticker.

3.2.2. Hand Shape

The shape of the hand, most often called manual configuration, constitutes one of the phonological parameters of the signs (Stokoe, 1960). Hand shapes can also be used lexically, with the status of Size and Shape Specifiers, or Proforms described below:

Size and Shape Specifier. At the boundary between lexicon and morphology, these specifiers are hand shapes used to depict the shape or size of objects to which the discourse refers (Moody, 1983; Cuxac, 2000; Millet, 2019). They may represent adjectives, i.e, they can be added to lexical signs, for example with *a small or big book*. In this case, the lexical sign "book" is executed first, then its size is delimited. They may also be *lexicalized*, like the signs "bowl" or "glass". In that case, the shapes of the hand refer to that of a bowl (Figure 1 a.) or a glass (Figure 1 d.).

Static Proform. Static proforms consist in using a specific hand shape representing an object (e.g. a flat hand for a table) or a person (e.g. a raised index finger for a standing person or a curved one for a sitting person). They are intuitive, efficient and powerful syntactic tools, particularly adapted to play the role of pronouns in sentences, by referencing lexical items, or to describe situations of relative spatial positioning. For example a pencil can be represented by the index finger (Figure 1 c.).

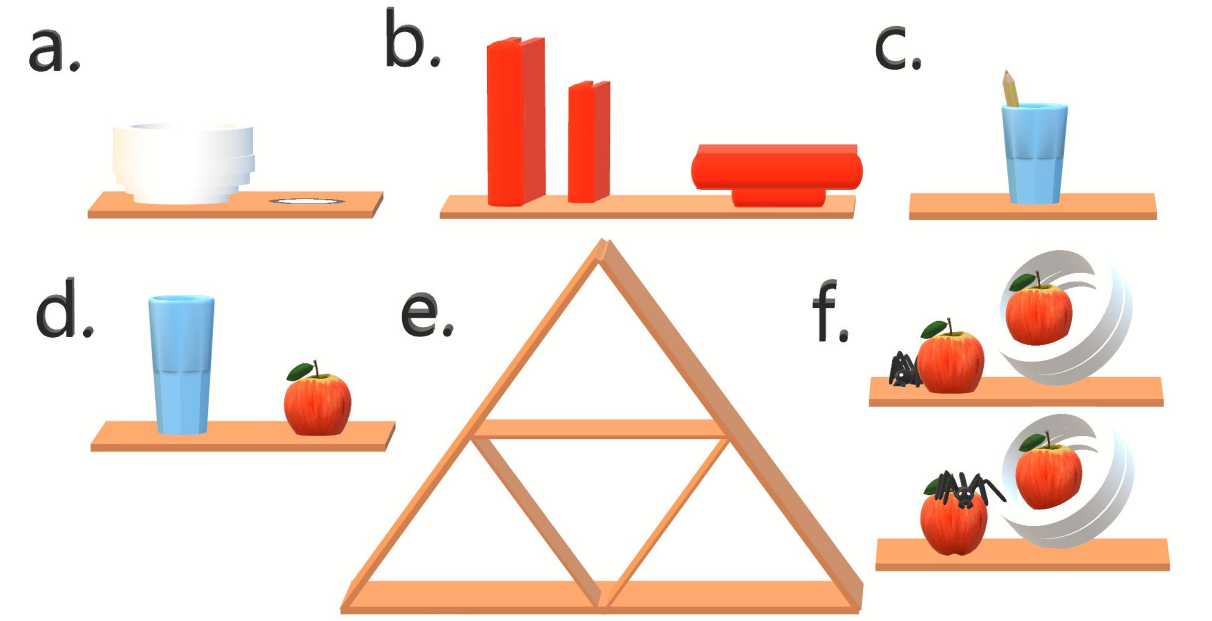


Figure 1: Different situations requiring specific spatial referencing mechanisms: a. a book is next to a sticker on the shelf; b. two books are aligned and two are stacked; c. a pencil is inside a glass d. a glass is next to an apple; e. an assembly of triangle-shaped shelves; f. (top) an apple (with a spider behind it) is next to a inclined bowl containing another apple; (bottom) the spider has climbed the first apple.

3.2.3. Multiple Positioning

We also consider in our corpus mechanisms of multiple arrangement of objects, which involve horizontal alignments (*sweeping*) and vertical stacks (*stacking*) (Millet, 2019). To describe these arrangements, hand movement is used, specifying the type of arrangement, while the shape of the hand keeps an iconic memory of the multiple objects. In this situation, hand shape may represent a manual proform if the object allows it (like a pencil or a plant) or a shape/size specifier (e.g., the hand shape C can represent cylindrical objects for describing buildings). This process refers either to a specific number of items or to an undefined number of objects (e.g., the description of a car in a traffic jam). Figure 1 b. shows a set of books aligned or stacked. In our corpus the number of items will be specified. Two situations are encountered:

Use a numeral sign. This corresponds to signing a numeral sign, such as ”3”, before signing the multiple entities, aligned or stacked.

Repeat aligned/stacked signs. This corresponds to signing in sequence several signs separately along a motion trajectory.

3.2.4. Relative Referencing

Relative placement is a 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 realized by static proforms that represent entities of the scene, thus ensuring the syntactic consistency of the sentence (Millet, 2019). Two kinds of situations are incorporated in our corpus:

Point and place. In this case, the signer points to a specific *Locus* and then signs the entities at this location (spatialization). For example, in Figure 1 d., a glass is signed on the shelve location and an apple is signed near it.

Relative Proform. Static proforms can be used to describe entities in the scene and their relative placement. For example, in Figure 1 c., two static proforms are used, a pencil (pr-pencil) and a glass (pr-glass). The passive hand represents the glass, while the active hand describes the pencil inside, with the possibility to orient precisely the pencil into the glass.

3.2.5. Geometric Shape Description

To describe complex geometric forms, in which subsets of objects are assembled or positioned in relation to each other, several iconic mechanisms, taken among the previous ones, can be used. Size specifiers, often combined with shape specifiers (Millet, 2019; Filhol and McDonald, 2020) are used to describe the volumes and dimensions of the shape. Moreover, relative referencing or proforms, with the passive hand as a reference for a predefined part of the shape and the active hand that describes, can be used to incrementally describe the complex shape. In Figure 1 e., the passive hand will be flat or angled to show where the current plank is placed in relation to the corners/segments of the pre-described part.

3.2.6. Hand Movement and Dynamic Proform

Dynamic proforms refer to the attribution of a verbal value to the lexical sign (Millet, 2019). Most of the time, this mechanism integrates moving proforms: animated entities, depicted as dynamic proforms, are thus used to describe the movements of these entities (e.g., displacement along a trajectory, jumping), as well as the expressive quality of the movements. The challenge here will be to recognize both the proform and its trajectory. Our corpus includes two types of situations:

Locus to Locus Movement. In this case, the motion of the moving entities is described by a trajectory anchored in the signing space. It starts from *Locus a.* and ends at *Locus b.*, or goes through a sequence of *Locus*.

Referential Proform. In this case, the passive hand describes an object around which the dynamic proform moves. For example, in Figure 1 f., the passive hand represents an apple while the active hand describes the movement of the spider climbing on it.

3.3. Content of the LSF-SHELVES Corpus

The corpus consists in sentences that describe situations in a signing scene by exploiting the spatial referencing mechanisms mentioned above. It mainly contains descriptions about bookshelves (from one to three shelves) with static entities (a bowl, a sticker, etc.) or dynamic ones (a spider) located on these shelves, with varying contexts. The lexicon of this corpus is composed of signs that can adopt specifiers, either shape or size, or that can be represented as proforms (static or dynamic). We specify in Table 1 whether the shape/size specifier, when it exists, is nominal (i.e. the specifier and the qualitative information contained in it represent a single lexical sign) or adjectival (i.e. a sign following a lexical sign), because this may impact the number of signs needed: one for a nominal shape/size specifier (e.g., [Bowl]), two for an adjectival shape/size specifier (e.g., [Book]). We also specify the ”prototypical aspect” (Millet, 2019) of the sign (iconic or not, neutral area or location on the body). An iconic sign in a neutral space can thus keep the same structural form when a verb is applied to it (e.g., the frog that jumps, or the pencil that is placed in the glass will both re-use features of the lexical term such as the handshape to modulate the verb), when it cannot otherwise.

Note that to this lexicon we have added the signed numbers ”1”, ”2”, ”3”, ”4” and ”5”. The corpus consists in 60 scenes divided in 6 levels of 10 scenes, each level adding complexity to the previous one and using the linguistic mechanisms described in 3.2.

Level 1 presents a single object on a shelf, sometimes with a sticker. It focuses on absolute spatial referencing (placement and spatialization for the main object and pointing for the sticker). The objects presented to the signers are boxes, bowls and glasses. Each can be specified in a nominal way with respect to its size and shape.

Level 2 contains alignments or stacks, and also describes repetitions of objects. Adjectival and nominal signs are also represented in this level. For example, the sign [Book], which has a size specifier but is adjectival, is used in descriptions of multiple objects.

Level 3 has several objects randomly arranged on two or three shelves. With non-aligned/stacked objects, it is necessary to use the process of relative referencing. Some of the new objects introduced in this level (e.g., pencil, plants) are easier to manipulate with proforms that are described in their static form.

Level 4 introduces geometric shapes using a variety of shelves assemblies and shapes (e.g., triangular, round). There is only one apple on the structure.

Level 5 uses the same type of shapes as Level 4, but with randomly arranged objects. Although it does not incorporate new linguistic mechanisms, it is the first level to mix relative referencing applied both to a lexical item or a shape.

Level 6 is similar to Level 3, with an animal moving through the scene. The animals considered, a spider and a frog, use different proforms and adopt different movements in the scene.

4. Acquisition of the Data

4.1. Capturing Motion Data with Kinect Azure

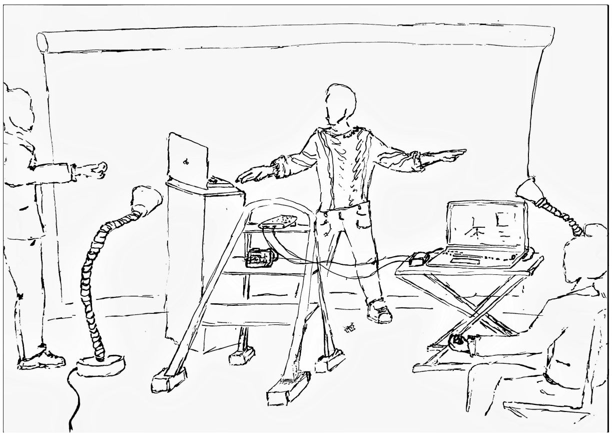


Figure 2: Captation installation scheme

For convenience, the captured motion data were stored as animated 3D skeletons. Indeed, this corpus is to be used in a machine learning context that deals with data represented as vector sequences. To that end, a suitable solution was to work with the *MoCap* technology which requires specific equipment.

In order to simplify the capture process for our volunteers, we preferred to use a portable system with one infrared camera such as the *Kinect Azure* (Sych et al., 2020) rather than a high-resolution system. The device relies on two NIR Laser diodes enabling near and wide field-of-view (FoV) depth modes, automatic per pixel gain selection enabling wide dynamic range, thus allowing near and far objects to be captured, global shutter for acquisition in sunlight and a multi-phase depth calculation method. In addition, it integrates an inner

Item	Shape/Size Specifier	Proform	Prototypical Aspect	Variations
Plank	Length and shape (nominal)	No	Iconic in neutral space	1
Box	Length, width and height (nominal)	No	Iconic in neutral space	2
Book	Length, width and height (adjectival)	No	Iconic in neutral space	1
Bowl	Diameter and height (nominal)	No	Iconic in neutral space	1
Glass	Height (nominal)	No	Iconic in neutral space	1
Ruler	Length (adjectival)	Yes	Non-iconic in neutral space	1
Pencil	No	Yes	Iconic in neutral space	1
Lamp	No	No	Iconic in neutral space	3
Apple	No	No	Iconic anchored on the body	1
Teddy	No	Yes	Iconic anchored on the body	1
Plant	No	Yes	Iconic in neutral space	1
Sticker	Shape (adjectival)	No	Iconic in neutral space	1
Spider	No	Yes (animated)	Iconic in neutral space	1
Frog	No	Yes (animated)	Iconic in neutral space	1

Table 1: List of lexical signs presented in our corpus. Column 1 is the name of the item, 2 indicates if they can admit a shape or size specifier, and if they are nominal or adjectival. Column 3 indicates if they admit a proform and if it is a static or animated one. Column 4 describes the prototypical value of the sign describing the item and column 6 the number of sign variation this item can be represented by.

software relying on a deep learning model (neural network) to extract a 3D skeleton from heat map (estimators). The resulting file format is *BVH* whose characteristics were defined by the acquisition software.

As can be seen on Figure 2, the resulting set up is very light and the constraints are relatively low. The layout may vary slightly from one capture session to another since they are often performed at the volunteer's place. The volunteer stands in front of the camera and preferably with a uniform background. Neither the brightness of the room nor the location of the volunteer seemed to have any impact on the quality of the data. In most setups, simple office lamps are used. The distance from the volunteer to the camera varies from 3 to 6 meters.

Each capture session lasts between 1 to 2 hours and resulted in an average of 30 minutes of raw data. The volunteer is therefore alone in front of the camera. He signs successively the 60 scenes that are provided to him. The instructions are presented on a computer dedicated to it, remotely controlled by the operator, an assistant or the volunteer himself. The operator also checks the capture process on the computer hosting the *Kinect* acquisition software.

4.2. Software

The *Motion Up* software (Le Naour, 2021) was used to extract the 3D skeleton from the *Kinect Azure* stream. This software handles a skeleton acquisition of 32 joints while correcting data to limit errors and allow further use of the data. This correction consists in recalculating the angular values to allow the continuity of the angular trajectories over time.

Unlike the *MoCap* acquisition room where markers can be added on the hands, the *Kinect Azure* is unable to recognize hands' configurations and orientations. To tackle this issue, we use a video recording of the captured movements in order to keep track of these hand poses over time. This video stream is then used in a post-processing step to edit the captured sequence and add animated handshapes with corresponding motions.

4.3. Signers and Elicitation

Despite the sanitary context, we were able to gather 15 volunteers with proficiency level in LSF varying from A2 to C2. Since the corpus is essentially based on spatial description, it presents referencing elements that are all mastered from level A2 in LSF. We enrolled a heterogeneous panel of volunteers. This heterogeneity is partly based on the level of fluency, but also on the volunteer's environment (geographical and sociocultural) as well as on the frequency and context of LSF use. It introduces some variability in the production of signed sentences, e.g., some signers prefer to point to a Locus while others directly sign objects at this specific location, while the level of fluency impacts the precision of descriptions. Thus, the panel is composed of volunteers aged from 20 to 70, practicing LSF either in their daily life, in the context of their studies, or in their professional environment, with varying frequencies of use. The volunteers' profiles are fairly distributed, as can be seen in Figure 3.

The signers did not know the detailed content of the corpus before the motion capture session. They signed 60 scenes divided into 6 levels (10 scenes per level), and had the opportunity to practice several times before signing each scene. During the recording session, they interacted with a Power Point presentation slide show, either directly or with the help of an assistant. Each slide showed a static scene consisting of a set of shelves with objects on them. This facilitated access to multiple views, by translating the 2D figure along horizontal and vertical axes, or by adding written aids (object's name and relative information such as ”small”).

Each signer was free to interpret the scenes and describe them as he/she wished, only a few instructions were given at the beginning. Thus, he/she was not

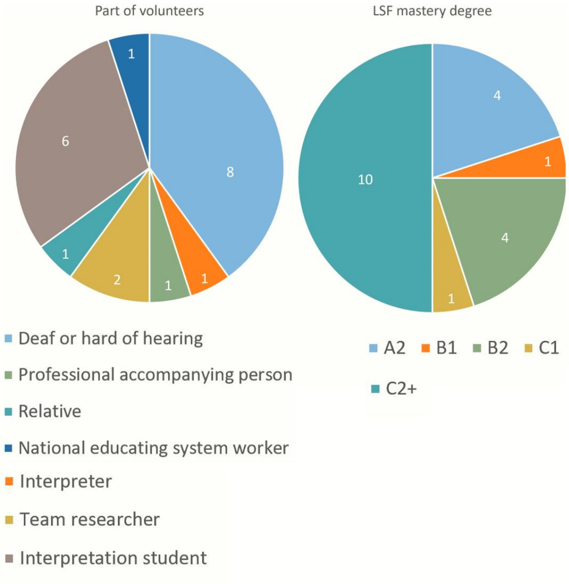


Figure 3: Left diagram represents the typical volunteer's profile, right diagram the proficiency according to European language proficiency categorization

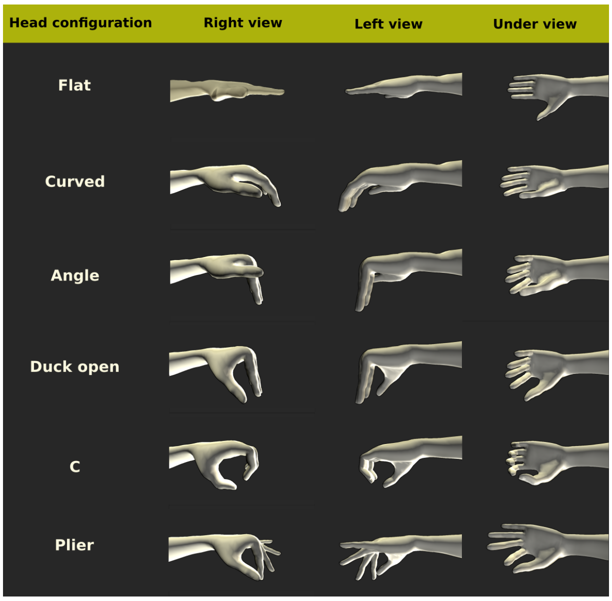


Figure 4: Examples of 6 out of the 32 handshapes that are used in our corpus

given specific instructions on how to sign the spatial referencing, but was constrained to use only the signs presented in an introductory sequence (e.g., the frog entity has several distinct signs, only one was used in the corpus). In addition, he/she was instructed not to use additional signs, especially to support descriptions (e.g., the "small" sign, colors) or to give details about placement (e.g., the signs "in the center" or "at the end").

Therefore, the instruction protocol described above highlights the flexibility of the French sign language introduced in our corpus, including: 1) flexibility in grammatical choice; 2) flexibility in order description choice; 3) flexibility of the language shortcuts that were introduced, such as two-handed lexical signs that can be executed with one hand (while this type of dynamics is not part of the grammar of LSF, it is common for fluent signers to use them). An additional layer of variability was also introduced in our corpus by the possibility of hand inversion (the active hand being alternately right or left).

5. Post-processing

As the motion data extracted from the *Kinect Azure* system is rather reliable, the main aspect of data post-processing consists in adding the missing hands (not provided by the device), by positioning and orienting them at the right spatio-temporal location in the movement sequence. To solve this problem, the *Motion Up* software (Le Naour, 2021) proposes a process in two successive phases. First, by adding hand poses through the use of a handshape keyframing technique. Second, by changing the orientation of the wrists. We describe below in more details these two post-processing phases, as well as the filtering system.

5.1. Handshape Keyframing

A hand pose database is first created and implemented within the *Motion Up* software. As seen in section 4.3, lexical signs were selected to limit the number of poses present in the corpus. This choice should increase the consistency of the future machine learning process. Our corpus contains 32 handshapes that are partially presented in Figure 4.

Using this handshape database, the process modifies BVH files obtained during the capture session by adding nodes representing the hands. Poses are directly selected in the software interface. Then, pose transitions are managed by the software and no additional processing is required in the resulting file.

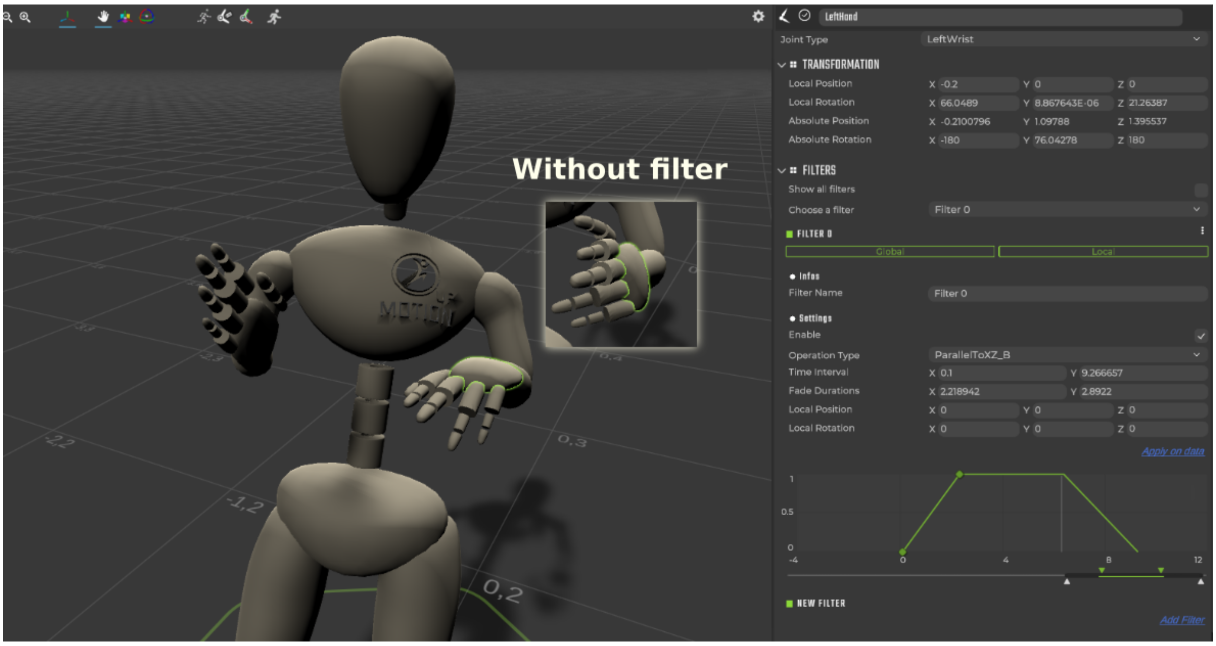
5.2. Adding Wrist Orientations

In the second phase of the post-processing, the question of the wrists orientation is particularly sensitive. Unlike the hand poses, wrist orientations are present in the raw data, but they carry errors in their angle trajectories, and need to be replaced manually. Thus, rotations over the three axis need to be calculated separately and manually overwritten.

5.3. Filtering Data

The *Motion Up* software integrates a filtering system that allows to apply transformations on the angular rotation trajectories (e.g., transform, add) and to modify them (e.g., value, duration, fading modulation).

The originality of this software lies in some additional filters that have been created to handle specific cases that are very useful for processing LS movements. In the case of wrist orientation, we mainly distinguish two



Joint Type	LeftWrist		
Local Position	X -0.2	Y 0	Z 0
Local Rotation	X 66.0489	Y 8.867643E-06	Z 21.26387
Absolute Position	X -0.2100796	Y 1.09788	Z 1.395537
Absolute Rotation	X -180	Y 76.04278	Z 180
Show all filters	□		
Choose a filter	Filter 0		
Filter Name	Filter 0		
Enable	☑		
Operation Type	ParallelToXZ_B		
Time Interval	X 0.1	Y 9.266657	
Fade Durations	X 2.218942	Y 2.8922	
Local Position	X 0	Y 0	Z 0
Local Rotation	X 0	Y 0	Z 0

Figure 5: Example of a filter modifying wrist rotation trajectory to obtain data where hand palm is parallel to the floor. On the right, some editing operations such as 3D position or rotation transformations, filtering, or interpolation

kinds of such editing techniques applied on rotations: 1) palm parallel/perpendicular to the floor or the other hand, and hand in the continuation of the forearm 2) palm parallel to the floor and palm pointing at some locus in front of the signer

These processes may force a body segment (the hand palm for wrist correction) to remain parallel to a specified plan such as illustrated in Figure 5. In that particular example, plan defined by the hand palm and the one defined by the floor will maintain parallelism during the time set in the filter settings.

Filters also allow to synchronize two trajectories so that they may mirror each other (applied for example on both arms' wrists trajectories). Finally, the *Motion Up* software integrates procedural models that make possible to perform inverse kinematics computations. This can be especially helpful in making the palm of the hand point to a specific location in space. All these filters make it possible to accelerate the post-processing time imposed by the absence of hand motion capture.

In its current state our corpus has not been through post-processing yet. The process has been tested and validated, post-processing of the data is in progress, and annotation should begin shortly.

6. Conclusion

We presented the *LSF-SHELVES* corpus, a new MoCap corpus of French Sign Language for recognition of iconic and spatial referencing mechanisms. LSF-SHELVES has been designed to present grammatical parameters such as the location of entities relatively to spatial references or other entities. It contains very few lexical signs and a wide variety of different spatial contexts. This corpus has been collected with the use of a novel low-cost protocol. This protocol being lightweight, we produced a consistent and precise data set that presents sufficient variability to be used in future recognition tasks using recent deep learning methods. Once the ongoing post-processing and annotation tasks will be achieved, this corpus will be open to the community. The following step will consist in producing an intermediate representation of the corpus, based on spatial referencing graphs. This formalism will be an important step toward the representation of descriptive sentences and a baseline for future automatic recognition of grammatical sign language mechanisms.

Furthermore, this research may lead to the implementation of a visual tool facilitating the understanding and learning of written grammar for deaf people.

7. Acknowledgments

We would like to thank M. Olivier Sainte-Claire from the Académie de Versailles, for his support and discussions on the corpus, as well as the charity association *Les Mains Diamant*, and the students of LSF interpretation master degree from university Paris VII, for their commitment on the motion capture.

8. References

Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Yuan, Q., and Thangali, A. (2008). The american sign language lexicon video dataset. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2008)*, pages 1–8.

Belissen, V., Braffort, A., and Gouiffès, M. (2020). Experimenting the Automatic Recognition of Non-Conventionalized Units in Sign Language. *Algorithms*, 13(12):310–345.

Benchiheub, M.-e.-F., Berret, B., and Braffort, A. (2016). Collecting and Analysing a Motion-Capture Corpus of French Sign Language. In *Workshop on the Representation and Processing of Sign Languages*, Portoroz, Slovenia, January.

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2021). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186.

Cihan Camgöz, N., Koller, O., Hadfield, S., and Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, pages 10020–10030.

Cooper, H., Holt, B., and Bowden, R., (2011). *Sign Language Recognition*, pages 539–562. Springer London, London.

Cui, R., Liu, H., and Zhang, C. (2017). Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 1610–1618.

Cuxac, C., Braffort, A., Choisier, A., Collet, C., Dalle, P., Fusellier, I., Jirou, G., Lejeune, F., Lenseigne, B., Monteillard, N., et al. (2002). Corpus ls-colin.

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

De Beuzeville, L., Johnston, T., and Schembri, A. C. (2009). The use of space with indicating verbs in auslan: A corpus-based investigation. *Sign Language & Linguistics*, 12(1):53–82.

DGS-Korpus. (2021). Hamburg university. `https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html`.

Dilsizian, M., Tang, Z., Metaxas, D., Huenerfauth, M., and Neidle, C. (2016). The importance of 3D motion trajectories for computer-based sign recognition. In *Proceeding of the Workshop on the Representation and Processing of Sign Languages: Corpus Mining, part of the International Language Resources and Evaluation Conference (LREC 2016)*, pages 53–58.

Ding, L. and Martinez, A. (2009). Modelling and recognition of the linguistic components in american sign language. *Image and vision computing*, 27:1826–1844, 11.

Efthimiou, E., Fotinea, S.-E., Hanke, T., Glauert, J., Bowden, R., Braffort, A., Collet, C., Maragos, P., and Goudenove, F. (2010). Dicta-sign: Sign language recognition, generation and modelling with application in deaf communication. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages:Corpora and Sign Language Technologies, part of the Language Resources and Evaluation Conference (LREC 2010)*, pages 80–84, 05.

Filhol, M. and McDonald, J. C. (2020). The synthesis of complex shape deployments in sign language. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives (LREC 2020)*, pages 61–68, Marseille, France. European Language Resources Association (ELRA).

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

Forster, J., Koller, O., Oberdörfer, C., Gweth, Y., and Ney, H. (2018). Improving continuous sign language recognition: Speech recognition techniques and system design.

Gibet, S. (2018). Building French Sign Language Motion Capture Corpora for Signing Avatars. In *Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, part of the International Conference on Language Resources and Evaluation (LREC)*, pages 53–58.

Huang, J., Zhou, W., Zhang, Q., Li, H., and Li, W. (2018). Video-based sign language recognition without temporal segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr.

Isaacs, J. and Foo, S. (2004). Hand pose estimation for american sign language recognition. In *Thirty-Sixth Southeastern Symposium on System Theory, 2004. Proceedings of the*, pages 132–136.

Junwei, H., George, A., and Alistair, S. (2009). Modelling and segmenting subunits for sign language recognition based on hand motion analysis. *Pattern Recognit. Lett.*, 30(6):623–633.

Koller, O., Ney, H., and Bowden, R. (2016). Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pages 3793–3802.

Konstantinidis, D., Dimitropoulos, K., and Daras, P. (2018). A deep learning approach for analyzing video and skeletal features in sign language recognition. In *Proceedings of the IEEE International Conference on Imaging Systems and Techniques (IST 2018)*, pages 1–6.

Kopf, M., Schulder, M., and Hanke, T. (2021). Overview of Datasets for the Sign Languages of Europe, July.

Le Naour, T. (2021). Motion up. `http://www.motion-up.com/`. Accessed: 2022-01-03.

Li, D., Rodriguez, C., Yu, X., and Li, H. (2020a). Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469.

Li, D., Yu, X., Xu, C., Petersson, L., and Li, H. (2020b). Transferring cross-domain knowledge for video sign language recognition.

Lu, S., Metaxas, D., Samaras, D., and Oliensis, J. (2003). Using multiple cues for hand tracking and model refinement. In *Proceedings of the IEEE/CVF Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003)*, volume 2, pages II–443.

Metaxas, D., Dilsizian, M., and Neidle, C. (2018). Linguistically-driven framework for computationally efficient and scalable sign recognition. In *Proceedings of The International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Millet, A. (2006). Le jeu syntaxique des proformes et des espaces dans la cohésion narrative en lsf. *Glottopol*, 7:96–111.

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

Moody, B. (1983). *La langue des signes, Tome 1 : Histoire et grammaire (French) [French Sign Language - First Volume: History and grammar]*. International Visual Theatre (IVT).

Naert, L., Larboulette, C., and Gibet, S. (2020). LSF-ANIMAL: A motion capture corpus in French Sign Language designed for the animation of signing avatars. In *Prodeedings of the International Conference on Language Resources and Evaluation (LREC 2020)*, pages 6008–6017, Marseille, France, May. European Language Resources Association.

Ojala, S., Salakoski, T., and Aaltonen, O. (2009). Coarticulation in sign and speech. In *NODALIDA 2009 workshop Multimodal Communication–from Human Behaviour to Computational Models*, page 21.

Ormel, E., Crasborn, O., and van der Kooij, E. (2013). Coarticulation of hand height in sign language of the netherlands is affected by contact type. *Journal of Phonetics*, 41:156–171.

Ormel, E., Crasborn, O., Kootstra, G., and de Meijer, A. (2017). Coarticulation of handshape in sign language of the netherlands: A corpus study. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 8(1).

Pu, J., Zhou, W., Zhang, J., and Li, H. (2016). Sign language recognition based on trajectory modeling with hmms. In *Proceedings, Part I, of the 22nd International Conference on MultiMedia Modeling - Volume 9516*, MMM 2016, page 686–697, Berlin, Heidelberg. Springer-Verlag.

Pu, J., Zhou, W., and Li, H. (2019). Iterative alignment network for continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*, pages 4160–4169.

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

Sych, T., Meadows, P., and Allen, B. (2020). Azure Kinect DK depth camera. `https://docs.microsoft.com/en-us/azure/kinect-dk/depth-camera`.

Vogler, C. and Metaxas, D. (2004). Handshapes and movements: Multiple-channel asl recognition. In *Lecture Notes in Computer Science*, pages 247–258. Springer.

Waldron, M. and Kim, S. (1995). Isolated asl sign recognition system for deaf persons. *IEEE Transactions on Rehabilitation Engineering*, 3(3):261–271.

Yuntao, C. and Weng, J. (2000). Appearance-based hand sign recognition from intensity image sequences. *Computer Vision and Image Understanding*, 78(2):157–176.