



# LSF-ANIMAL: A Motion Capture Corpus in French Sign Language Designed for the Animation of Signing Avatars

Lucie Naert, Caroline Larboulette, Sylvie Gibet

## ► To cite this version:

Lucie Naert, Caroline Larboulette, Sylvie Gibet. LSF-ANIMAL: A Motion Capture Corpus in French Sign Language Designed for the Animation of Signing Avatars. LREC 2020, May 2020, Marseille, France. hal-03005767

**HAL Id: hal-03005767**

**<https://hal.science/hal-03005767>**

Submitted on 1 Apr 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# *LSF-ANIMAL*: A Motion Capture Corpus in French Sign Language Designed for the Animation of Signing Avatars

Lucie Naert, Caroline Larboulette, Sylvie Gibet

IRISA Lab, Université Bretagne Sud

Vannes, France

{lucie.naert, caroline.larboulette, sylvie.gibet}@univ-ubs.fr

lsf.irisa.fr

## Abstract

Signing avatars allow deaf people to access information in their preferred language using an interactive visualization of the sign language spatio-temporal content. However, avatars are often procedurally animated, resulting in robotic and unnatural movements, which are therefore rejected by the community for which they are intended. To overcome this lack of authenticity, solutions in which the avatar is animated from motion capture data are promising. Yet, the initial data set drastically limits the range of signs that the avatar can produce. Therefore, it can be interesting to enrich the initial corpus with new content by editing the captured motions. For this purpose, we collected the *LSF-ANIMAL* corpus, a French Sign Language (LSF) corpus composed of captured isolated signs and full sentences that can be used both to study LSF features and to generate new signs and utterances. This paper presents the precise definition and content of this corpus, technical considerations relative to the motion capture process (including the marker set definition), the post-processing steps required to obtain data in a standard motion format and the annotation scheme used to label the data. The quality of the corpus with respect to intelligibility, accuracy and realism is perceptually evaluated by 41 participants including native LSF signers.

**Keywords:** Corpus, Motion Capture, French Sign Language, Marker Set, Annotation, Perceptual Evaluation

## 1. Introduction

Virtual characters, or *avatars*, are a promising solution to convey information in Sign Language (SL). Avatars make it possible to preserve the signer’s anonymity and to gain in interactivity: the speed, appearance and 3D point of view of the signed sequence can easily be changed by the user. In addition, unlike videos, the signed production can potentially be edited to create new signed content.

However, in order to be accepted by the Deaf, the avatars must be fully animated with natural, realistic and meaningful motions. Those motions can be (i) generated from **hand-crafted keyframes** (Braffort et al., 2007; McDonald et al., 2016), (ii) generated **procedurally** (Kennaway, 2003; Nunnari et al., 2018), or (iii) generated with **data-driven techniques** through concatenation of pre-captured motions (Gibet et al., 2011) or machine learning processes (Brock and Nakadai, 2018). In the first case, the quality of the produced animation can be very high but the process of manually creating the keyframes is fastidious and time-consuming. In the second case, the generated motions are often robotic and unrealistic but any sign can be produced as long as it can be described using a SL representation such as SigML (Kennaway, 2006), or Azee (Fihol and Falquet, 2017). In the third case, the avatar’s motions come from human data and are thus more natural and smooth but the variety of the signs that can be synthesized is limited by the initial motion capture (MoCap) data set.

As we considered the acceptance of the avatar to be a prime issue, we chose to use pre-captured motions but, to overcome the limitations of this approach, we seek to enrich the initial corpus by generating new signs through the recombination of motion segments on the different body channels such as the hand configuration or the wrist orientation. The initial corpus must be defined precisely so that the signs and motions present in the corpus can be used to study the sign formation mechanisms and to generate new signs. Besides,

the MoCap corpus must be fully annotated in order to be used both to analyze and to synthesize signs and utterances. Moreover, for the synthesis results to be correct and accepted by the Deaf, the initial signs and movements present in the MoCap data set must satisfy requirements in terms of precision and realism that can only be assessed by a qualitative evaluation of the data.

This paper presents the *LSF-ANIMAL* corpus, a motion captured corpus in French Sign Language (LSF) focused on the manual aspects of SL. Section 2. reviews existing video and motion capture data sets for different sign languages. Section 3. describes the objectives and the content of the *LSF-ANIMAL* corpus. Section 4. presents the data acquisition process and the detailed marker set used. Section 5. details the post-processing and annotation steps. Finally, Section 6. presents the design and results of a perceptual evaluation of the quality of the data set.

## 2. Related Work

Sign languages are visual-gestural languages. Given the lack of exhaustive and widely accepted written representations of SL, only video cameras or motion capture technologies can provide sign languages recordings accurate enough to be used for analysis or synthesis.

Video recordings constitute the most common source of data. The subjects are filmed from one or more points of view, data is stored using a standard video format and is annotated *a posteriori* following a pre-defined annotation template. Video corpora are often the base material for statistical studies to highlight a particular gestural phenomenon or to verify a given hypothesis on human motion. Various video databases have been designed by linguists in order to study a specific linguistic feature (e.g., directional verbs and signing space (De Beuzeville et al., 2009), coarticulation (Ormel et al., 2017), iconicity and role playing in the LS-COLIN corpus (Cuxac et al., 2002), classifier

predicates in narratives/stories (Millet, 2006)); to compare different SL and different dialects of the same SL (Schembri and Johnston, 2004; Hanke et al., 2017); or simply to archive and preserve signed utterances (Crasborn and Zwitterlood, 2008). Other SL resources have been designed to perform natural language processing (Efthimiou and Fotinea, 2007) or automatic recognition (Camgöz et al., 2016; Ebling et al., 2018) tasks. Videos can be used for the study and analysis of SL but do not provide reusable data for SL synthesis. The cost of a video recording session is quite cheap as a single video camera commercially available may be sufficient. However, a video recording alone eliminates the third dimension of space. Pose estimators such as *OpenPose* (Cao et al., 2018) or additional cameras are needed to compute the depth information during the realization of the signs. Besides, video recordings rarely possess a spatial resolution and a frame rate high enough to allow a precise data segmentation and analysis.

Motion Capture (MoCap) technologies offer a higher spatial and temporal resolution than 2D video cameras in exchange for the need of a greater technical expertise and a rigorous post-processing. Data resulting from a motion capture session can be used for SL analysis: precise quantitative motion descriptors can be computed from the 3D data to confirm or reject existing linguistic hypotheses or motion laws.

As the 3D positions of the human skeleton joints can be inferred from the MoCap data, avatars can also be animated from the captured data which constitutes a major advantage compared to video data. So, while a vast majority of video corpora are designed for linguistic analyses and computer vision tasks, motion capture corpora purposes are evenly distributed between **analysis** (e.g., coarticulation analysis (Ormel et al., 2013) or kinematic analysis (Benchiheub et al., 2016)) and **data-driven synthesis** (e.g., concatenative synthesis for French Sign Language (Gibet, 2018) or deep neural network approach for Japanese Sign Language (Brock and Nakadai, 2018)). In American Sign Language, the CUNY corpus (Lu and Huenerfauth, 2012) is used to study linguistic mechanisms which are exploited to build a linguistic representation that can be used to animate an avatar.

However, even though their number is steadily growing, MoCap databases for SL studies are still rare and a very small portion of them are made available. Besides, MoCap databases are small compared to video databases: they rarely exceed one hour of data and contain the sign utterances of few different signers (often only one signer), while a video footage can last hundreds of hours and gather the data of various persons. As a consequence, each of the existing MoCap corpus has been designed for a specific purpose. In our case, we aim to synthesize realistic LSF signs and utterances by analyzing, editing and/or recombining the captured motion segments on the different manual channels. For example, we plan to create a sign *A* by extracting and possibly modifying a feature (e.g., the hand configuration) of a sign *B* (see example on Figure 1). We use the *LSF-ANIMAL* corpus to study some linguistic mechanisms, the motion kinematic features on the different manual channels and the synchronisation between the channels, and as

raw material to create new signs and utterances.



Figure 1 – Two different LSF signs using the same motion but different configurations: *snail* with the 'Y' configuration (left) and *slug* with the 'U' configuration. The small offset between the two arms is due to the retargeting process and the presence of markers on the signer's arms.

### 3. Corpus Definition

Because sign languages do not contain a finite number of signs due to mechanisms such as classifier-predicates, and because capturing data is costly, covering all the signs in all the possible contexts is not a viable design solution. Defining a corpus specifically suited to the studied phenomenon and the objectives to be achieved is thus more relevant.

#### 3.1. Objectives

The objective of our corpus is dual. On the one hand, it constitutes the material to be analyzed in order to highlight motion laws, invariants and LSF phenomena. In this case, the data can be considered as ground truth and is used to make observations and to evaluate our synthesis results. On the other hand, the data becomes the synthesis material. We aim to generate new, natural and realistic LSF utterances based on the observations of the ground truth, and editing of the captured motions. This analysis/synthesis complementarity is paramount for our research work and the corpus is designed to handle this duality. More precisely, we wish to study and synthesize three manual LSF parameters (Stokoe, 1960) (sometimes referred to as "phonological elements" (Johnston and De Beuzeville, 2010)):

**1 - Hand configurations (HC) of LSF.** Keeping our synthesis goals in mind, we aim to study two phenomena: the transition from one configuration to another and the synchronisation of the HC with respect to the other channels such as hand orientation and placement. The corpus must then incorporate these HC in various linguistic constructions: in signs containing a change in the manual configuration – e.g., the sign [SALON] (*living room*) begins with an 'O' and ends with a 'C' configuration –, as well as in full utterances in which the chosen HC appears in a natural and contextualized way. Moreover, for synthesis purposes, the isolated HC must be captured to serve as a basis to our synthesis system.

#### 2 - Placement of the two hands in the signing space.

The placement of the hands can designate both (i) the global area where the sign is produced which does not change during the sign production but which can change depending on grammatical inflections (Millet, 2019; Moody, 1983b), and (ii), at a lexical level, the discrete area or the specific coordinates where the hand is positioned at a precise time. Both

are interesting for our study. We need to capture instances of the same sign placed at different locations of space and signs in which the hands are not static and whose position vary. The capture of full utterances will naturally provide various placement features.

**3 - Hand movement.** Hand movements with different trajectories (straight line, circular, waves, etc.), and with zero or more repetitions must be incorporated into the database.

In addition, we wish to study **coarticulation** mechanisms in full utterances. In order to measure the impact of the sign  $N - 1$  and  $N + 1$  on the sign  $N$ , it is necessary to record natural LSF utterances with various combinations of the same signs.

Finally, we want the corpus to be used to test automatic annotation algorithms with different levels of granularity (e.g., gloss, hand configuration, placement). For this purpose, the presence of various types of data streams, such as isolated HC, isolated signs and full sentences, is beneficial.

### 3.2. Content of the LSF-ANIMAL Corpus

To meet the requirements detailed in the previous section, the *LSF-ANIMAL* corpus contains four subsets (see Table 1).

The first subset constitutes a list of the most common **hand configurations of LSF**. 41 hand configurations have been chosen with care by comparing five sources of different nature: (i) a LSF teacher, (ii) the hand configurations annotated in the Sign3D Corpus (Gibet et al., 2016), (iii) the International Visual Theatre book which is a reference for LSF grammar and vocabulary (Moody, 1983a), (iv) the research book of (Cuxac, 2000) and (v) a textbook to learn LSF (Amauger et al., 2013). This part contains all the letters used for fingerspelling in LSF<sup>1</sup> which can be used to spell some words and names, and 22 isolated configurations. All those configurations were executed with both hands.

The second subset is composed of **isolated signs**. Three types of signs have been chosen to address three types of needs. (i) 11 signs containing a change of hand configuration within the sign (like the sign [WEEK-END] in LSF which begins with the 'W' and ends with the 'E' hand configurations). Those signs can be used as examples and ground truth to synthesize the passage from one configuration to another and to study the coarticulation of the hand configuration channel within a sign. (ii) 9 question words (*where?*, *when?*, *what to do?*, *how?*, *how old?*, *what?*, *why?*, *who?*, *how much/many?*). Those signs are crucial in LSF: in addition to their function as interrogative pronouns, they are used to explain the context

of a situation in relative clauses. The sign *where?* can thus introduce the place of the action and the sign *why?* can mean "because" in an affirmative sentence. (iii) 47 names of animals (e.g., *ostrich*, *duck*, *whale*) and 25 animal descriptors (e.g., *mammals*, *feather*, *blue*). Each animal name was performed twice to ensure the presence of two almost identical signs in the resulting data. Having a choice between different instances of the same motion segment is beneficial to add realism in a synthesis context. Animal names present a great range of contextualized hand configurations. Those configurations are therefore performed in different locations in the signing space. Animal names and descriptors can also have recreational applications and can be used, in serious games, to teach the signs corresponding to animals to a French hearing or deaf population.

The third subset consists in 26 **descriptions of animals** in four categories (6 dogs, 5 cats, 11 birds, and 4 mammals with horns). The color, type of skin (fur, plumage, etc.), food preferences and habitat are described for each animal. Then, the animal is identified with a name (*seagull*, *black dog*, *cow*, etc.). This task was inspired by a LSF lesson for beginners. It provides a natural and authentic flow of LSF utterances. Transfers of person in which the signer impersonates the character that he/she is talking about (Sallandre and Cuxac, 2001) are numerous in such descriptions as the signer will naturally imitate the animal he is describing. In addition, the resulting production contains various different hand configurations, placements and types of motion.

At last, the fourth subset focuses on three **grammatical mechanisms** of LSF (and of SL in general): size and shape specifiers, pointing gestures and classifier predicates. The particularity of those three mechanisms is that only one manual feature (resp. movement amplitude, hand placement and hand configuration/movement) changes to modify the meaning of the sign/utterance<sup>2</sup>. This property is very interesting for the synthesis of new content by recombination of the body channels.

**1 - Size and shape specifiers** consist in using the standard sign of an object with a different amplitude of movement to accurately represent the size/shape of the object (Sallandre and Cuxac, 2001; Supalla, 1986). A big bone, for example, will be signed by doing the sign [OS] (*bone*) with a larger amplitude than for a normal bone. It is very interesting as only one feature changes (the amplitude of the motion) when the same object is described with different sizes. We chose to capture some examples of size specifiers based on the text of a popular fairy tale, namely *Goldilocks and the Three Bears* in which a young girl finds herself interacting with various objects of three different sizes.

**2 - Pointing gestures** are paramount in sign languages as they can be used to designate the subject(s) of an action (*You*, *This theater* or *This man over there*) or to associate virtual objects to 3D locations in the signing space. Those objects can then be referred to using a pointing gesture on

<sup>1</sup>Not each of the 26 letters of the fingerspelling alphabet is a hand configuration. The 'N' letter, for example, possesses the same configuration as the 'U' letter (index and middle fingers in the up position while the other fingers are folded). However, the alphabet alone represents 19 different configurations.

<sup>2</sup>The information conveyed by the face and gaze are also an important part of these structures but they are not the focus of this study.

| # | Task name              | Content                                                                                                  | Purposes                                                                                                                       | Duration per signer                   |
|---|------------------------|----------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------|---------------------------------------|
| 1 | Hand configurations    | Fingerspelling alphabet + 22 isolated configurations                                                     | Training set for automatic annotation and synthesis of the hand configurations.                                                | Signer 1: 2 min<br>Signer 2: 3 min 30 |
| 2 | Isolated Signs         | 11 signs with a change in the hand configuration + 9 question words + 47 animal names and 25 descriptors | Analysis of hand configuration transitions, presence of words reusable in various contexts, recreational/educational purposes. | Signer 1: 9 min<br>Signer 2: 5 min    |
| 3 | Continuous signing     | 26 descriptions of animals in 4 categories                                                               | Study of transfers of person, of coarticulation and of the impact of the manual parameters of LSF.                             | Signer 1: 20 min<br>Signer 2: 9 min   |
| 4 | Grammatical mechanisms | Size and shape specifiers + pointing gestures + classifier predicates                                    | Recombination of the manual parameters of LSF for synthesis purposes.                                                          | Signer 2: 10 min                      |

Table 1 – Content of the *LSF-ANIMAL* corpus.

the 3D location. To capture various pointing gestures, we designed a task in which the signer had to point with his index to various places on its body and in the signing space. Other types of pointing gestures exist, involving the gaze, the shoulder or torso movements but we limited this study to index pointing gestures following the definition of Blondel (Blondel, 2009).<sup>3</sup>

**3 - Classifier predicates** consist in using a particular hand configuration to represent an object (e.g., a flat hand for a car) or a person (the index finger raised for a standing person, bent for a sitting person, etc.) and a movement of the hand to show the movement performed by the object/person. For example, two flat hands moving forward with one hand behind the other one will depict two cars driving in a row. Classifier predicates are used to describe a scene more vividly and with fewer signs. In our case, classifier predicates have two interesting features: they take full advantage of hand configurations and show a wide range of movement trajectories that can be reused in different contexts. We chose to record 18 utterances describing different situations involving vehicles and pedestrians classifiers (moving forward, crossing each other, etc.).

## 4. Acquisition of the data

The capture must follow a strict protocol to collect clean and usable data. Technical considerations, the signers' profile and the elicitation protocol are presented hereafter.

### 4.1. Technical considerations

In addition to the definition of the corpus, it is necessary to prepare the capture room and to define a marker set in accordance with the task.

#### 4.1.1. Motion Capture Room

The capture of French Sign Language utterances can be performed in a limited space as the signer does not move during the linguistic production but it also brings important technical constraints (Courty and Gibet, 2010): (i) the need to accurately capture gestures with small but meaningful variations (the finger motions particularly), (ii) the temporal dynamics (velocity, acceleration, jerk) must be preserved which requires a high sampling rate and (iii) the whole body is involved in sign language production: facial expressions, gaze, torso motions and manual characteristics must be captured simultaneously.

For the capture, we used a Qualisys environment composed of sixteen infrared cameras (eight OQUS 400 and eight OQUS 700) and one video camera. To preserve the dynamics of the language, the capture was performed at a sampling rate of 200Hz.

#### 4.1.2. Marker Set

The choice of the position and size of the optical markers attached to the signer's body are very important to capture the linguistic production with precision. A trade-off must be made between the quality of the capture and the intrusiveness of the equipment. Markers with a large radius will be visualised more easily by the infrared cameras than markers with a smaller radius but will impede the signer's motion and will thus impact the quality of the resulting data. Bigger markers were placed on body locations which are not prone to collide with other body parts while smaller markers were attached to the fingers and the face.

We used 123 optical markers in total for our capture (see Table 2). For the body (without considering the hand and facial markers), the locations described in (Carreno, 2015) were chosen. It consists in putting two markers of large radius (12.7 mm) around each joint position (elbows, knees, wrists, ankles) and at other strategic places (sternum, back, feet, etc.) (see Figure 2, left).

Facial markers are paramount to capture the facial expressions which are meaningful in LSF utterances but also, and most of all considering our objectives, to capture the areas where the hand touches the face in some signs. To capture subtle deformations and to interfere as little as possible with the signer, those facial markers have a small diameter (4 mm). Given that the hands and their movements are at the center of our study, only 16 markers have been placed on the face; they are a subset of the facial marker set of (Reverdy et al., 2015) and form a coarse cartography of the face and its main elements (nose, forehead, mouth, cheek, chin) that serve to indicate the position of the hand with respect to the face.

A more thorough study was performed to determine the location of the hand markers. In order to capture accurately the complexity of the hand motion, it is necessary to use numerous small-sized markers. The performance of reduced hand marker sets (down to six markers on the hand) to produce natural motions were compared in (Hoyet et al., 2012). The authors of the article mainly sought to obtain a realistic motion for simple tasks. However, in addition to realism, sign languages require the avatar motions to be identical or very similar to the source motion. Besides, the location of the markers on the hand is very important to subsequently

<sup>3</sup>Note that, with our synthesis engine, we are capable of associating any hand configuration to those pointing gestures.

| Marker set | Nb of markers | Marker diameter          |
|------------|---------------|--------------------------|
| Head       | 16            | 12 × 4 mm<br>4 × 12.7 mm |
| Hands      | 26 (×2)       | 19 × 4 mm<br>7 × 6.5 mm  |
| Body       | 55            | 12.7 mm                  |

Table 2 – Our marker set: number, diameter and location of our 123 markers.

reconstruct the hand skeleton from the data. The right part of Figure 2 shows a skeleton of the hand. A marker was placed on each of the MCP joints, the closer to the bone as possible, and two markers were put on each PIP joints. One marker was added on the extremity of the second and third phalanges. This way, every first and second phalanges are defined by isosceles triangles. It gives an indication of the finger width and direction and simplifies the recognition of phalanges in post-processing work (in particular during the labeling of unidentified trajectories of markers).

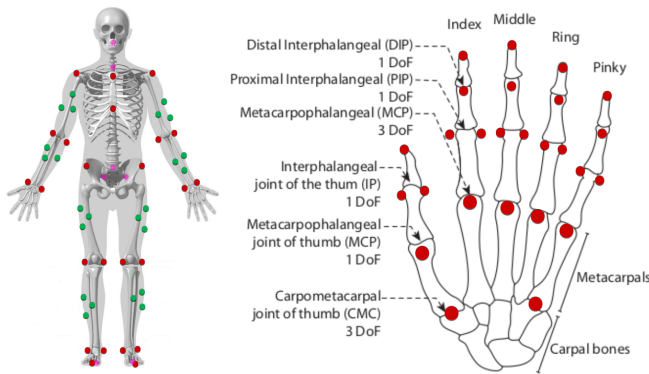


Figure 2 – Left: body marker set. The red markers are used for the definition of the segments and for the tracking whereas green markers are only used for tracking. Pink markers are situated on the back. Right: skeleton and marker set of the right hand.

## 4.2. Signers and Elicitation

The four subsets of the corpus were captured on two deaf LSF instructors fluent in written French (called Signer 1 and Signer 2 in Table 1). The instructions for each task were displayed on a screen in front of the signers. Before each new task, the instructions were clarified in LSF by a member of the lab. The signers knew in advance the global content of the corpus but they discovered the precise tasks during the capture session.

A trade-off was found between precisely controlling the corpus content and giving the signers enough freedom to have the most natural and realistic sign language production. To take this into account, three of the tasks were precisely controlled with instructions written explicitly while the signers were given some leeway in the third part in which he/she has to describe various animals. For this task, we needed the signers to be able to sign in the manner they see fit. We decided to give the signers an image of the animal to be described next to some words underlining the important information that the signers must provide in their description. No sentences were imposed. The signers could add as many details as they wanted.

The two capture sessions lasted 4h30 each. We obtained

around 1h of raw data in total ( $\approx 30$  min for each signer).

## 5. Post-processing

### 5.1. General Post-processing

We ensured that the markers were correctly identified by manually correcting the errors and labeling the markers that were not identified by the Qualisys software (Qualisys, 2019). On some frames, the occlusion of some optical markers led to the absence of motion data for those markers. The position of each occluded marker on each frame was reconstructed by interpolation following the work of (Le Naour et al., 2018), thus ensuring that the 123 markers were visible on each frame. MotionBuilder (Autodesk, 2018) was used to derive the position and orientation of the actual joints of a human skeleton from the positions of the markers. This led to the creation of motion files in FBX, a standard motion format. To visualize and evaluate our data, we have then rigged a character to the skeleton defined in the motion files. However, for our work which consists in studying and editing the MoCap data, it is also important to annotate the data.

### 5.2. Annotation

The data annotation process consists in associating to each frame of the captured data one or more labels describing the movement performed during this frame. Annotation consists of (i) dividing a continuous stream of movements into smaller segments and (ii) labeling these segments. These tagged segments will then be retrieved to be studied or to animate an avatar. As the final animation of the avatar depends on these labels, it is essential to have a precise and shared definition of the movement segments and labels.

The content of the *LSF-ANIMAL* corpus was annotated on the ELAN software (MaxPlanckInstitute, 2017) after the post-processing. Our annotation followed a structural scheme adapted to data-based synthesis, and was achieved at different levels. The glosses/signs were annotated manually by one person with knowledge of LSF on different annotation channels (left hand, right hand and both hands). To remove the main biases of the manual annotation of glosses, it was automatically refined using motion features (Naert et al., 2017). Moreover, given that the focus of our work is the study of the different manual parameters of sign languages, 12 other channels, corresponding to the configuration of each hand, the placement of the hands, the motion type and orientation were created. Hand configurations were automatically annotated using machine learning methods (Naert et al., 2018). The placement channels were also automatically annotated by computing distances between the hands and the body/facial markers. To this day, the orientation and motion channels are only partially manually annotated. Our corpus is therefore composed of two dependent data sets: the captured motions and the annotations. To manipulate motion segments, the annotation data set can be queried and the corresponding motions are retrieved from the MoCap data set (Gibet et al., 2011).

## 6. Perceptual Evaluation of the Corpus

As the initial corpus serves as the core material for the analysis and synthesis of movements, the quality of the

synthesis will depend on the quality of the corpus. It is thus necessary to assess whether the signs and motions of the corpus are accurate and realistic. We therefore evaluated the quality of the data present in the corpus using a perceptual evaluation on a subset of the corpus.

To this end, we formulated the following hypotheses :

$H_1$ : The captured data is intelligible.

$H_2$ : The captured data is accurate.

$H_3$ : The captured motions are realistic.

$H_4$ : No information is lost when post-processing the data.

## 6.1. Design of the Evaluation

To validate or reject our hypotheses, we created videos by varying independent parameters. We then randomly showed the videos to the participants and asked them to recognize the video content and grade the realism and accuracy on a 5-point Likert scale.

### 6.1.1. Evaluated videos

The *LSF-ANIMAL* corpus is a compound of isolated signs, signs in context and utterances. The physical descriptions of various animals constitute a large part of the corpus. It is difficult and irrelevant to segment these descriptions into discrete signs because, since they are not standard, they would not have a meaning without their context. We therefore decided to evaluate the corpus by presenting two types of sequences: (i) isolated standard signs (animal names) and (ii) whole utterances corresponding to animal descriptions. Participants were asked to find the meaning and to evaluate the accuracy and realism of the sequences. The isolated signs were chosen in order to test if small differences in the manual parameters of LSF were visible. For example, we chose the signs *bird*, *goose* and *duck* which are identical in movement and placement but different for the hand configuration. In a similar way, *tiger* and *zebra* are identical except for the hand placement (on the head for *tiger* and on the torso for *zebra*).

In addition, in order to be able to validate hypothesis  $H_4$ , it was necessary to show different types of data representations at different stages of the processing. Three representations were selected: (A) the points in space linked by segments given by the Qualisys software after identification of the trajectories, (B) the skeleton with the position of the joints calculated by MotionBuilder, and (C) the avatar controlled by the skeleton (see Figure 3).

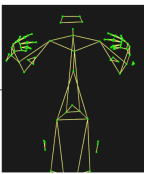
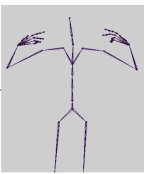
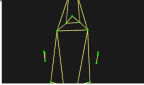
| Representation<br>Sequence | (A) Qualisys:<br>3D points                                                          | (B) Skeleton                                                                        | (C) Avatar                                                                          |
|----------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| Isolated Signs             |  |  |  |
| Descriptions               |  |  |  |

Figure 3 – The different combinations of sequences and types of representations.

We chose to evaluate 18 sequences : 9 isolated signs and 9 descriptions (see Table 3) multiplied by 3 types of representations (54 videos in total). To avoid overloading the participants, we separated the evaluation into 3 sub-evaluations of

18 videos. Each sub-evaluation was composed of two parts (Part 1: isolated signs, Part 2: descriptions) and each part contained 9 videos of 9 different sequences. No sequence was evaluated more than once by the same participant, not even with two different representations but they all saw the 3 representations. The videos could be played back as many times as the participants wanted. One additional video per part was used as a training session.

| #        | 1: Isolated signs | 2: Descriptions   |
|----------|-------------------|-------------------|
| 1        | Bird              | Labradoodle (dog) |
| 2        | Eagle             | Dachshund (dog)   |
| 3        | Duck              | Eagle             |
| 4        | Goose             | Rooster           |
| 5        | Tiger             | Grey cat          |
| 6        | Zebra             | Duckling          |
| 7        | Horse             | Persian cat       |
| 8        | Frog              | Red cat           |
| 9        | Mouse             | Black kitten      |
| Training | Cat               | St Bernard (dog)  |

Table 3 – The 18 signs and described animals (plus the 2 training sequences).

### 6.1.2. Questions

Three questions per video were asked to the participants. A question testing the intelligibility of the sign or description ( $H_1$ ), a question on accuracy ( $H_2$ ) and a question on the naturalness of the realization ( $H_3$ ). All questions and their possible answers were signed in LSF and the resulting videos were subtitled in French.

At first, only the question testing intelligibility was visible. In the case of a sign, the question asked was: "What sign was made?" The answer was in the form of a drop-down list of animal names containing about fifty animals including the correct answer, with, at the end, two additional lines: "I did not recognize the animal" and "The animal I recognized is not present in the list" (chance level of 2%). In the case of a description, the question asked was: "Which image best matches the description that has been made?". Nine answers were proposed: the correct image, 7 images close to the description but that did not match exactly the description + the answer "No image matches the description" (chance level of 11%). The suggestions were close enough so that the answer was not obvious.

When the participant had validated his/her answer to this first question, the correct answer appears (e.g., "It was the sign DOG") and the following questions are made visible. The second question concerns the accuracy and precision of the sign or description: "Do you think that the sign/description was done correctly?". The third question concerns the naturalness of the movement: "Do you think that the sign/description in LSF is natural/realistic/spontaneous (does it seems to be the movement of a real person)?". In both cases, possible responses are presented on a Likert scale ranging from 1 (most negative) to 5 (most positive).

## 6.2. Results

We released the questionnaire online and collected the results from 41 participants, 12 men and 29 women with an

average age of 39.46 years old (min = 19, max = 72). Among the participants, 22 were hearing people ("*entendant*"), 3 were hearing-impaired ("*malentendant*"), 2 had become deaf during his/her lifetime ("*devenu sourd*") and 14 were deaf since birth ("*sourd de naissance*"). In addition, the participants were asked to assess their level of French Sign Language (*beginners*: 13 participants, *quite good*: 6, *good*: 5, *very good*: 8 or *natives*: 9)<sup>4</sup>.

### 6.2.1. Recognition Rate

The recognition rate with respect to the level of French Sign Language of the participants is shown in Figure 4. In a logical way, the *very good* and *native* signers achieve a better recognition rate for isolated signs than participants with a lower level of LSF.

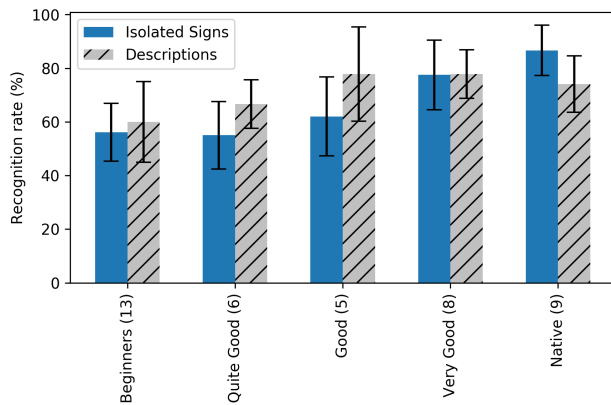


Figure 4 – Recognition rate of isolated signs (blue bars) and descriptions (hatched bars) with respect to the level of LSF of the participants (the number of people in each category is specified between parenthesis).

The significant difference between the recognition rate of isolated signs of the *beginners* and the *natives* ( $p$ -value of  $6.65e^{-05}$  with the unilateral Mann-Whitney test) shows the non-triviality of the task. More generally, people with a *good* level and below have a better recognition rate for descriptions than for isolated signs.

This can be due to several reasons: (i) a random response on the description part is more likely to be correct as there are fewer possible answers in the description part than in the isolated signs part, (ii) unlike animal names which are isolated standard signs, the descriptions of animals are contextualized and iconic sequences: even people not knowledgeable in LSF can have an idea of the animal described just with the impersonation of the signer (e.g., the behaviour of the signer when describing the dachshund's walk or the labradoodle's curly hairs), (iii) participants can learn signs as they watch the videos: the vocabulary to describe the type of fur, for example, is repeated on different descriptions while, in the case of signs, learning was useless because no two signs were identical.

Isolated signs and descriptions were correctly identified with an average recognition rate of 76.068% by the *good*,

*very good* and *native* LSF signers. Figure 5 shows their recognition rate per sequence. We can see that 9 out of 18 sequences have a recognition rate higher than 80% even though we intentionally chose similar signs (e.g., *bird*, *duck*, *eagle* and *goose*).

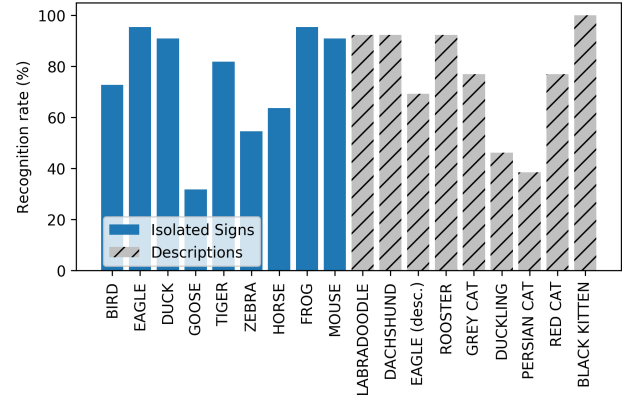


Figure 5 – Recognition rate per sequence for the "good", "very good" and "native" LSF signers (22 participants).

For the isolated signs, *goose* has the lowest results. 23% of the *good*, *very good* and *native* LSF signers mistook *goose* for *turkey* which shows the same hand configuration at a slightly different location (on the mouth for the goose and the nose for the turkey). The participants who replied *turkey* had the avatar representation, the only representation with the head visible: the placement may be slightly off on the avatar.

For the descriptions part, *persian cat* and *duckling* were not associated with the correct picture in a majority of cases. For both of them, the picture that was chosen instead represented an animal with a color specified in the description (an orange cat instead of an orange-eyed persian cat and a brown duck instead of a brown duckling).

However, participants can make mistakes and it is therefore important to analyze the answers to the following questions concerning the accuracy and the realism of the signs.

### 6.2.2. Accuracy and Realism

We considered that only the *good*, *very good* and *native* LSF signers were legitimate to answer the accuracy and realism questions. So we exclusively took into account their answers in this section. Table 4 shows the mean accuracy and realism scores while Figure 6 details the answers of the participants.

|                 | Isolated Signs  | Descriptions    |
|-----------------|-----------------|-----------------|
| <b>Accuracy</b> | 3.545/5 (0.703) | 3.701/5 (0.584) |
| <b>Realism</b>  | 3.667/5 (0.398) | 3.957/5 (0.425) |

Table 4 – Accuracy and realism average scores of the *good*, *very good* and *native* participants. The standard deviation is specified inside parenthesis with respect to the scores per sequence.

With more than 3.5 out of 5, we considered the realism of the movement to be acceptable. The reconstruction of the movement thus provides realistic human motions.

As for the accuracy of the signs, some signs, such as *horse*, were not recognized, not because of a problem in our processing but due to the original movement. Indeed, *horse* is

<sup>4</sup> Among the 14 *deaf since birth*, only 9 indicated a *native* level of LSF.

usually done with both hands but, in our database, it is done with only one hand which can explain the poor results. As its form is not the standard form of the *horse* sign, it will be discarded from the database. Apart from those signs, the median of the results in accuracy are high (equal or above 4) for a majority of sequences (12 out of 18 sequences).

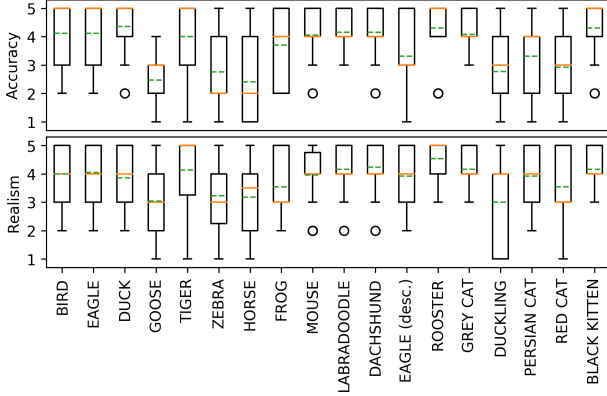


Figure 6 – Accuracy and realism score per sequence (the median is the orange line, the mean is the dotted green line, the whiskers go to 1.5 multiplied by the interquartile range). For example, *bird*, *eagle* and *duck* have an accuracy of 5 according to half of the participants.

Except for some sequences that should be removed from the corpus, we consider that  $H_2$  and  $H_3$  are verified which means that the data can be used for synthesis work.

### 6.2.3. Impact of the Type of Representation

To verify the  $H_4$  hypothesis, the accuracy and realism scores were grouped by type of representation (Qualisys, Skeleton or Avatar, see Figure 3) for participants with a LSF level greater than or equal to *good*. Each participant rated between 9 and 18 sequences depending on whether or not he/she did the second part (among the 22 participants with a level greater than or equal to *good*, 13 responded to the two parts). We therefore gathered 105 ( $22 \times 3 + 13 \times 3$ ) realism ratings per representation (same for accuracy, see Figure 7).

As the data do not follow a normal distribution, we used the Kruskal-Wallis test for non-parametric data to determine if the type of representation had an impact on the ratings. Whether for accuracy or realism, the results of the statistical test do not allow us to rule out the  $H_4$  hypothesis (for accuracy:  $p\text{-value} = 0.10$  and, for realism:  $p\text{-value} = 0.65$ ). We performed unilateral Mann-Whitney tests for the accuracy results and obtained a result that was close to being significant between the skeleton and the avatar representation ( $p\text{-value} = 0.022$ ).

Ideally, we could have benefited from a higher number of results from LSF experts but, as it stands, we can consider that the type of representation had no significant impact on the ratings. The quality of the data was preserved in terms of realism from the raw MoCap data (Qualisys representation) to the animated skinned avatar. For the accuracy, there might be a small loss between the skeleton and the avatar representations but the current results do not allow us to reject the  $H_4$  hypothesis.

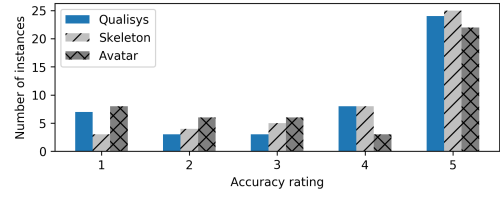


Figure 7 – Number and nature of the accuracy ratings per type of representation for the *good*, *very good* and *native* LSF signers.

### 6.2.4. Comments

At the end of the questionnaire, we allowed participants to express their feelings in a free text space. Out of the 41 participants, 9 commented on the lack of facial expressions, 6 expressed their preference for the skinned avatar, justifying it by the presence of placement information of the hands with respect to the face and by the fact that the avatar had a human appearance, 3 said they preferred the skeleton for the precision of the gestures, 3 others preferred the Qualisys representation for the same reason and 3 said they were enthusiastic about the precision and fluidity of the gestures. One participant expressed his surprise in understanding the two avatars without heads.

## 7. Conclusion

We presented the *LSF-ANIMAL* corpus, a new MoCap corpus of French Sign Language for sign language analysis and synthesis applications. The captured data has been post-processed so that it can directly be used to animate virtual signers. This task is still on-going and, at this time, the corpus has not yet been made available to the community. The corpus is composed of four subsets, each of them containing various signs and grammatical mechanisms to meet different synthesis objectives. Some manual parameters including hand configuration, placement, movement and orientation were captured and annotated. Therefore, our corpus, that includes semantically meaningful data, can be used for concatenative synthesis but can also be enriched by combining the different motion segments present in the data set and/or by editing the motion signal in order to create new content not limited to animal names and descriptions. And, given the exhaustive annotation scheme, this corpus can be used in other applications including LSF analysis. Besides, except for some specific signs, the results of the perceptual evaluation show that the movements were considered accurate and realistic by the participants and that the post-processing of the MoCap data did not impact the quality of the data in a significant way.

In the future, we hope to assess the precision of the motion capture process with respect to the manual parameters of LSF by taking advantage of the similarities between some animals (for example *bird*, *duck* and *goose*, whose only difference is the number of fingers used in the hand configurations to express the beak of the animals).

The perceptual evaluation of the corpus presented in this paper will serve as a baseline for future evaluations of the synthesis work.

## 8. Acknowledgments

We want to thank M. Irdel and C. Gendreau-Touchais for their commitment in the capture and evaluation of our work.

## 9. References

- Amauger, F., Bertin, F., Gonzalez, S., Tsopgni, P., and Vanbrugghe, A. (2013). Langue des signes Française - A1 (French) [French Sign Language - A1]. Langue des signes Française. Belin.
- Autodesk. (2018). Motionbuilder. <https://www.autodesk.com/products/motionbuilder/overview>. Accessed: 2019-04-19.
- Benchiheb, 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.
- Blondel, M. (2009). Acquisition bilingue lsf-français: L'enfant qui grandit avec deux langues et dans deux modalités. Acquisition et interaction en langue étrangère, (Aile... Lia 1):169–194.
- Braffort, A., Bolot, L., Filhol, M., and Verrecchia, C. (2007). Démonstrations d'elsi, la signeuse virtuelle du limsi (French) [Demo of elsi, the virtual signer of the limsi]. In colloque Traitement Automatique des Langues des Signes, Atelier Traitement Automatique des Langues des Signes.
- Brock, H. and Nakadai, K. (2018). Deep jslc: A multimodal corpus collection for data-driven generation of japanese sign language expressions. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).
- Camgöz, N. C., Kindiroğlu, A. A., Karabüklü, S., Keleş, M., Özsoy, A. S., and Akarun, L. (2016). Bosphorusign: a turkish sign language recognition corpus in health and finance domains. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 1383–1388.
- Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2018). OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In arXiv preprint arXiv:1812.08008.
- Carreno, P. (2015). Marker-set specification for magician's motion capture database recording. Personnel document.
- Courty, N. and Gibet, S. (2010). Why is the Creation of a Virtual Signer Challenging Computer Animation ? In Motion in Games 2010, LNCS, pages 1–11, Netherlands.
- Crasborn, O. A. and Zwitserlood, I. (2008). The corpus ngt: an online corpus for professionals and laymen. In 3rd Workshop on the Representation and Processing of Sign Languages. Paris ELRA.
- Cuxac, C., Braffort, A., Choisier, A., Collet, C., Dalle, P., Fusellier, I., Jirou, G., Lejeune, F., Lenseigne, B., Monteillard, N., et al. (2002). Corpus lsc-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.
- Ebling, S., Camgöz, N. C., Braem, P. B., Tissi, K., Sidler-Miserez, S., Stoll, S., Hadfield, S., Haug, T., Bowden, R., Tornay, S., et al. (2018). Smile swiss german sign language dataset. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).
- Efthimiou, E. and Fotinea, S.-E. (2007). Gslc: Creation and annotation of a greek sign language corpus for hci. In HCI.
- Filhol, M. and Falquet, G. (2017). Synthesising sign language from semantics, approaching "from the target and back". arXiv preprint arXiv:1707.08041.
- Gibet, S., Courty, N., Duarte, K., and Le Naour, T. (2011). The signcom system for data-driven animation of interactive virtual signers: Methodology and evaluation. Transactions on Interactive Intelligent Systems.
- Gibet, S., Lefebvre-Albaret, F., Hamon, L., Brun, R., and Turki, A. (2016). Interactive editing in french sign language dedicated to virtual signers: requirements and challenges. Universal Access in the Information Society, 15(4):525–539.
- 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, LREC 2018, Miyazaki, Japan, May.
- Hanke, T., Konrad, R., Langer, G., Müller, A., and Wähl, S. (2017). Detecting regional and age variation in a growing corpus of dgs.
- Hoyet, L., Ryall, K., McDonnell, R., and O'Sullivan, C. (2012). Sleight of hand: perception of finger motion from reduced marker sets. In Proceedings of the ACM SIGGRAPH symposium on interactive 3D graphics and games, pages 79–86. ACM.
- Johnston, T. and De Beuzeville, L. (2010). Auslan corpus annotation guidelines. Centre for Language Sciences, Department of Linguistics, Macquarie University.
- Kennaway, J. R. (2003). Experience with, and requirements for, a gesture description language for synthetic animation. In Proc. of Gesture Workshop 2003, LNCS, Genova, Italy, February.
- Kennaway, R. (2006). Avatar-independent scripting for real-time gesture animation. arXiv preprint arXiv:1502.02961.
- Le Naour, T., Courty, N., and Gibet, S. (2018). Kinematic driven by distances.
- Lu, P. and Huenerfauth, M. (2012). Cuny american sign language motion-capture corpus: first release. In Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, The 8th International Conference on Language Resources and Evaluation (LREC 2012).
- MaxPlanckInstitute. (2017). Elan v.4.9.4. <http://tla.mpi.nl/tools/tla-tools/elan/>. accessed 29 January 2019.
- McDonald, J., Wolfe, R., Schnepf, J., Hochgesang, J., Jamrozik, D. G., Stumbo, M., Berke, L., Bialek, M., and Thomas, F. (2016). An automated technique for real-time production of lifelike animations of american sign

- language. Universal Access in the Information Society, 15(4):551–566.
- 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. (1983a). La langue des signes, Tome 1 : Histoire et grammaire (French) [French Sign Language - First Volume: History and grammar]. International Visual Theatre (IVT).
- Moody, B. (1983b). 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. (2017). Coarticulation analysis for sign language synthesis. In International Conference on Universal Access in Human-Computer Interaction, pages 55–75. Springer.
- Naert, L., Reverdy, C., Larboulette, C., and Gibet, S. (2018). Per channel automatic annotation of sign language motion capture data. In Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, LREC 2018.
- Nunnari, F., Filhol, M., and Heloir, A. (2018). Animating azeze descriptions using off-the-shelf ik solvers. In 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community (SignLang 2018), 11th edition of the Language Resources and Evaluation Conference (LREC 2018), pages 7–12.
- 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).
- Qualysis. (2019). Qtm. <https://www.qualisys.com/>. Accessed: 2019-04-19.
- Reverdy, C., Gibet, S., and Larboulette, C. (2015). Optimal marker set for motion capture of dynamical facial expressions. In Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games, pages 31–36. ACM.
- Sallandre, M.-A. and Cuxac, C. (2001). Iconicity in sign language: A theoretical and methodological point of view. volume 2298, pages 173–180, 04.
- Schembri, A. and Johnston, T. (2004). Sociolinguistic variation in auslan (australian sign language): A research project in progress. Deaf Worlds, 20(1):S78–S90.
- Stokoe, W. C. (1960). Sign language structure: An outline of the visual communication systems of the american deaf. Studies in Linguistics, Occasional Papers, 8.
- Supalla, T. (1986). The classifier system in american sign language. Noun classes and categorization, pages 181–214.