Collecting and Analysing a Motion-Capture Corpus of French Sign Language

Mohamed-El-Fatah Benchiheub[1,2], Bastien Berret[2], Annelies Braffort[1]
[1]LIMSI-CNRS, University of Paris-Saclay, [2]CIAMS, Univ. Paris-Sud, University of Paris-Saclay
[1]Campus d'Orsay, bât 508 F-91405 Orsay cx, France , [2]Campus d'Orsay, bât 335 F-91405 Orsay cx, France
mohamed-el-fatah.benchiheub@u-psud.fr, bastien.berret@u-psud.fr, annelies.braffort@limsi.fr

Abstract

This paper presents a 3D corpus of motion capture data on French Sign Language (LSF), which is the first one available for the scientific community for pluridisciplinary studies. The paper also exhibits the usefulness of performing kinematic analysis on the corpus. The goal of the analysis is to acquire informative and quantitative knowledge for the purpose of better understanding and modelling LSF movements. Several LSF native signers are involved in the project. They were asked to describe 25 pictures in a spontaneous way while the 3D position of various body parts was recorded. Data processing includes identifying the markers, interpolating the information of missing frames, and importing the data to an annotation software to segment and classify the signs. Finally, we present the results of an analysis performed to characterize information-bearing parameters and use them in a data mining and modelling perspective.

Keywords: French Sign Language, Motion Capture, Mocap, Animation, Annotation, Movement Analysis.

1. Introduction

Sign languages (SLs) are languages used to communicate with and among the Deaf communities. They are natural languages based on visuo-gestural modalities. Recent advances in computer graphics and animation have allowed the possibility to create and display 3D content in SL, by using a virtual signer (or signing avatar), i.e. a 3D character expressing itself in SL. This method allows the broadcasting of messages to Deaf people in an anonymous and modular way. However, generating 3D models based on actual knowledge of SL kinematics is still a challenge for computer scientists.

French Sign Language (LSF), as many other SLs, is still little described, particularly for what concerns the movement of articulators, and the existing models or representations in computer science are very simplified. Most of the studies in SL processing are interested in modelling linguistic properties, but few are interested in understanding the kinematics or dynamics of the movement itself and how it improves the comprehensibility of the generated signing. The rare ones have been applied on video corpora that do not allow estimating accurately and reliably velocities and accelerations (Segouat and Braffort, 2009; Lefebvre-Albaret, 2010).

Getting a better account of SL motion data thus requires novel resources. Recording 3D kinematics will allow designing more accurate models and improving knowledge in all scientific disciplines related to SL. However, the availability and the accessibility of the necessary technologies, which is scarce and expensive, make 3D corpora still rare especially for LSF.

Existing studies based upon such 3D corpora showed that they are of great value for all applications: generation, analyse of the movements (kinematic and dynamic) as well as linguistic analysis. For example, an American Sign Language (ASL) corpus has been used to compare animations generated by motion capture (mocap) and by generation algorithms. It was found that the animation based on mocap data generates movements that are more natural (Lu and Huenerfauth, 2010). Another study using LSF mocap data has been dedicated to automatic segmentation of the hand movement based on principal component analysis (Héloir et al., 2006). This method proved to be effective to solve high-level segmentation. 3D corpora are also used for linguistic analysis. For example, a study focused on identifying the type of verb (Telic and atelic), which seem to be distinguishable on the basis of speed and acceleration parameters on ASL corpus (Malaia et al., 2008).

There exists 3D corpora for LSF (Duarte and Gibet, 2010), but either they are not available or they do not meet the requirements for multidisciplinary research as we envision it, which is animation replay with a virtual signer, 3D data analysis of both body and facial movements, and linguistic annotation.

For these reasons, we started to create APlus, a 3D corpus of LSF available to the scientific community for multidisciplinary studies[1]. Our paper presents the steps of the data recording, data processing (labelling, gap-filling), and annotation. We also demonstrate how we perform and may exploit kinematic analysis on 3D data.

2. Content of the corpus

This paper describes the first part of the corpus, which represent about one hour of data. Six LSF native signers were involved in this part. They present various socio-linguistic profiles and signing styles, in order to have some insights on inter-signer variability.

Signers were asked to describe pictures in a spontaneous way. Each signer had a look at each picture during a few minutes before beginning the recording session. The elicitation material consisted of a set of 25 pictures showing many objects with peculiar geometrical properties (e.g. horizontal or vertical arrangements etc.) as in Figure 1. For the subject, the task thus consisted of describing successively the images.

The second part of the corpus, including various tasks is not described in this paper. More details can be found in

[1]More details on corpus characteristics here: https://tals.limsi.fr/corpus

Figure 1: Example of the described pictures

(Braffort et al., 2015).

3. Data recording

Motion capture is the process of recording the movement of objects or people. The recorded mocap data is transformed into a digital format for further processing and analysis or mapped on a digital model in 3D software. The recording provides a numerical coordinate matrix that can be used as a source of data for analysing the movements of the body parts from a kinematic perspective.

All the recordings have taken place in our studio in the Complexité, Innovation and Activités Motrices and Sportives laboratory (CIAMS) at the University of Paris-Sud, France. The CIAMS laboratory focuses on the study of motor control from biomechanical, neurophysiological and psychological perspectives. The studio hosts a 10 camera optical motion capture system (OptiTrack S250e). The frame rate of the cameras is 250 Hz, which is a sufficient resolution for our purpose. It allows managing a correct amount of markers with various sizes, attached to the body but also to the face, where miniaturized markers are needed. Given the resolution of the cameras, we have designed a setup with 40 markers of various sizes allowing to track the motion of the limbs but also movements occurring on the face (eyebrow, eye lead, cheeks and mouth movements). However, our system does not allow for accurate tracking of all the fingers. Only coarse information is given on finger movements. In addition to the mocap cameras, we have also used a digital video camera that provides a classical video to be used in the annotation software.

The very first step of this work was to design the best setup for the camera and marker locations. For that, we have conducted evaluations such that we can record sufficient details of the human performance for our various needs: animation replay with virtual signer, 3D data analysis, and linguistic annotation. One of the most important questions in mocap data recording is marker locations: where to attach the markers on the body, and why? This issue is important because the location of markers affects their visibility in the system: covered markers are not recorded. Marker locations are also important from the point of view of potential post-processing steps such as transforming the three-dimensional marker data into joint or segment representations. Furthermore, markers that are placed inappropriately might make it difficult for the signer to properly articulate signs. Finally, marker location must allow us to track as much as possible all the useful movements from a linguistic point of view.

Figure 2 shows the setup of the forty markers that we have used. There are 4 markers on top of the head: 2 in the front and 2 in the back. The torso contains 7 markers: 4 on the upper part (sternum, clavicle, two on the back (C7: Spinous process of the 7th cervical vertebrae, and T10: Spinous process of the 10th thoracic vertebrae), the other 3 markers on the pelvis. Each arm has 5 markers placed on the main joint positions (shoulder, inner and outer elbow, wrist ulnar and radial) and one on the triceps. There are also 2 markers on each palm. A set of 13 markers is used for the face: eyebrows, eyelids, cheeks, chin and mouth (below, above, left and right).

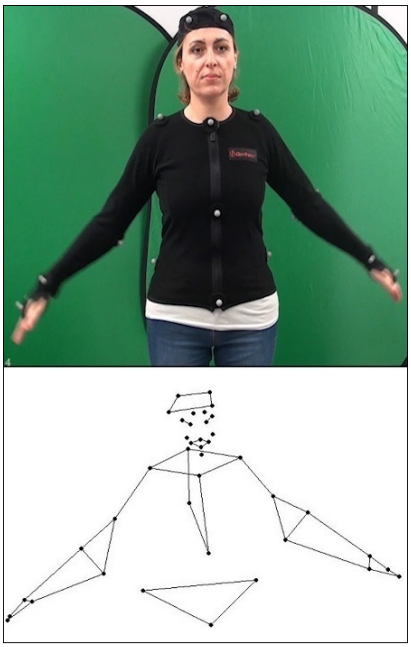

Figure 2: Up: markers attached on a subject, Down: markers connected by segments

The positions in our configuration were chosen so that the markers are maximally visible and identifiable by the system, and so that they capture the main global movements of the hands, arms, upper torso, and head. The location of the 6 markers on each arm was chosen in a way to be able to reconstruct the orientation (joint angles) of the 2 segments of the arm (upper arm and forearm). The rule is that there must be at least 3 markers on a rigid body to define its 3D orientation. We have put markers on the pelvis to differentiate the movements of lower part of the torso and those

of the upper part. We have done tests with markers on the fingers but with the 10 camera system, finger markers overlap between them. For that reason, we have used 2 markers on each hand palm that allow us to have at least movement and rotation of the hands. The markers on the face allow us to have the eyebrow movements, winks, movement of the cheeks and of the mouth.
In comparison with other recent mocap studies, the total number of markers in our setup is fairly comprehensive: (Jantunen et al., 2012) used 20 markers (7 on each arm and hand, 4 on the head , and 2 on the upper torso). (Tyrone et al., 2010) used 30 markers (7 on each arm, 7 on the head, and 9 on the torso) and (Duarte and Gibet, 2010) whose additional goal is to use the data to create animated signing, avatars employed 98 markers (43 facial markers, 43 body markers, and 6 on each hand).
Figure 3 shows our optimal camera setup. We have used 4 cameras facing the signer and at the same height of the signer's head, 2 cameras on each side, and 2 cameras behind the signer. The cameras ahead allow a very good capture of the face markers, the cameras on the sides and behind allowing to capture the markers of the arms and head, the cameras behind are sufficient to capture the 2 markers placed on the back (C7 and T10). The digital video camera is placed in front of the signer. This setup allowed us a very good capture with minimum losses and overlapping of markers during recording.

Figure 3: Setup of the ten-camera optical motion capture as well as the HD video camera in our studio

In comparison with other recent mocap studies, (Jantunen et al., 2012) used eight-camera optical motion capture system (Qualisys ProReflex MCU120), while (Lu and Huenerfauth, 2012) used the Animazoo IGS-190 system to capture the movement of the arms and torso, with Intersense IS-900 to capture the movement of the head together with the two Immersion Cyber Gloves and eye tracker to capture the hands and eye movements respectively. (Duarte and Gibet, 2010) used twelve-camera optical motion capture system (Vicon MX).
We have included in our tests the use of a Tobii eye tracker. This is a device that incorporates illumination, sensors and processing to track eye movements and gaze point. This device allows to record gaze direction. This kind of device is not satisfactory because it hides the eyebrow and eye-lids movements. A better device remains to be found in order to include eye gaze in our data.

4. Data processing

Once raw data are recorded, there are several essential steps that must be done before the data can be exported and exploited. Due to the possible occlusion between the various parts of the body, and because the markers are not identified and may appear identical (marker swapping), a post-processing is needed to clean up the data.
The use of a high number of markers (40 for this corpus) has a drawback, which is the amount of gaps in the data as well as the overlapping between the markers which are close, or which will be close during the signing. This disadvantage may be overcome by the use of a larger number of mocap cameras (here we used 10 cameras). At least 2 cameras must see a marker at each frame for its instantaneous three-dimensional location to be recovered. If we had a system with more cameras (18 or 20) we would not have the gaps in data or overlap between markers, as there would be seen at any time by at least 2 cameras.
The Optitrack Motive software[2] gives several setups of markers. Using these predefined setups, we could obtain directly the markers labelled at the end of the recording. Unfortunately, these setups do not take into consideration the face, so we did not use them, and we add a step of identification (labelling) of the markers. Each time a marker is lost it must be re-labelled.
When all markers are labelled, we move to the second stage, which consists of removing noise. Indeed, at the end of the recording, there are fake/phantom markers, which are due to noise or reflection during the recording. This step can be done automatically, by removing all remaining non-labelled markers.
The third step is the gap-filling (filling the missing frames). This step can be done in Matlab software after export of the data with specific toolboxes such as (Burger and Toiviainen, 2013).But these toolboxes give quite arbitrary results when the gap is too long. To solve this problem, an option of the Motive software has been used before the export, with several methods (linear interpolation, cubic interpolation or interpolation based on other markers). The interpolation relative to other markers (markers that are in the same segment and which are fixed to each other) is the best method for the gaps that are relatively long, the cubic interpolation was used for the gaps that are in circular movements, and linear interpolation was used for the gaps that are in linear movements.
The next step is to check if there is an overlap between markers (errors on the marker identification). There is an overlap when two markers are too close between them during recording sequences, and the system confuses the two markers and reverse their identity. There are two cases in the overlap:

- The first case, the identity of markers remains reversed after the overlap.

[2]http://www.optitrack.com/products/motive/

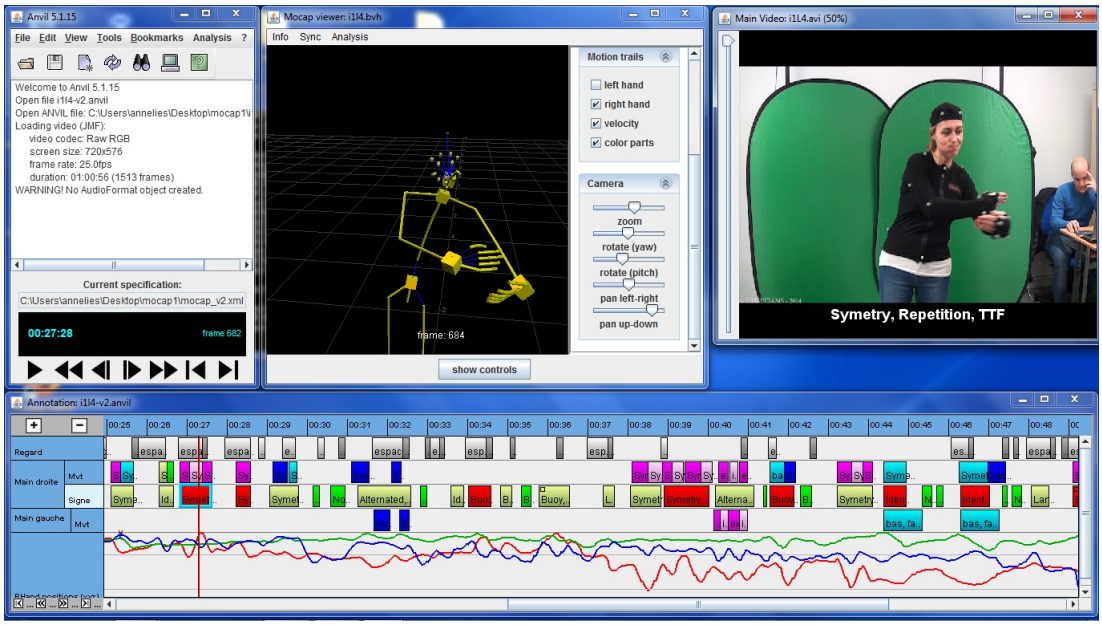

Figure 4: ANVIL screenshot shows the annotation using video, 3D skeleton and the mocap data

- The second case, the system reverses the identity of the markers only during the overlap, i.e. when the markers move away from each other their identities returns correct.

The verification is done marker by marker throughout the recording sequence. When there is an overlap between two markers, we delete the data of the two markers during the overlap. Then, as we said above, we have two cases. In the first case where the identity of markers remains reversed, we remove the labelling which is after the overlapping of these two markers, and we re-labelled them with the good identifications. In the second case the markers have the correct identification.

Now that we have markers with the good identifications, we fill the gaps that we made during the correction of the overlapping by using one of the three types of interpolation (defined above) depending on the case of movement and the sizes of the gaps.

At the end of these steps, and before exporting the data, we did an audit of the data by checking that all markers were labelled and that there were no gaps throughout the recording. This verification is done by running the animation in Motive software and looking at the colour of the markers. If they are all white during the animation, it means they are labelled throughout the recording. If a marker's colour changes from white to orange, this means that it is not labelled during these frames. To verify that there are no gaps in all markers, we verify that no marker is lost during all frames, but there is another easier way by selecting all markers and verify if there is no holes in displacement curves (X, Y and Z). When all these steps were achieved, the data were exported in c3d format with a frequency of 250 Hz.

The last step consists of making the data usable for the annotation software. ANVIL annotation software[3] was chosen because it can display the 3D data in addition to the video. For that, the c3d format was transformed into a bvh format by adding a skeleton hierarchy using the 3ds Max software.

[3] http://www.anvil-software.org/

5. Annotation

To make the data usable by linguists and also to analyse the movements, the bvh files and videos were imported to ANVIL annotation program. At this moment, the annotation is composed of three tracks, the first for gaze direction, the second for the type of movement (e.g. main direction), and the third for the linguistics annotation (see Figure 4). So far, the annotated movements are the linear ones in the three main axis up-down, medial-lateral, and anterior-posterior of both hands.

6. Data analysis

The 3D corpus enables an accurate quantitative analysis, allowing us to compute multiple parameters that characterize the movement: position, speed (mean velocity, peak of velocity), acceleration, angles between articulators, etc. As our ultimate aim is to develop models to generate LSF movements, we have first to identify the information-bearing parameters to reproduce in priority those critical parameters and get meaningful LSF. Indeed, it is currently difficult to expect that a model will reproduce all kinematic features of LSF given the complexity and the large number of degrees of freedom of the human body. Moreover, from a motor control viewpoint, the laws of motion used by signers when producing LSF movements are still poorly known, especially in comparison to non-LSF movements produced by other individuals. Intriguing and unresolved questions pertain to the existence of invariant and peculiar features in the kinematics of SL movement, and how they compare to non-SL movements.

We present here some preliminary results of our study related to velocity, and explain how this kind of corpus can be exploited for the study of motor control in SL and how simple process on 3D data can allow for automatic computation of metadata related to the signer.

6.1. Mean velocity: linked to the degree of control?

A parameter that has been analysed here is the mean velocity of the movement of the dominant arm, for lexical signs and depicting signs that describe the size and the shape of entities (SASS) which are very frequent in this description task. It was found that this parameter varies extensively between different subjects: For instance, the mean velocity in lexical signs and SASS respectively of a subject was around 0.51 m/s and 0.60 m/s with standard deviation of 0.3 m/s and 0.22 m/s across the entire session, while the mean velocity in lexical signs and SASS respectively of another subject was around 0.91 m/s and 0.97 m/s with standard deviation of 0.31 m/s and 0.23 m/s.
This drastic change of movement pace gives some hints about the underlying control laws used in LSF. In particular, it allows to assert that the mean velocity of movement of the dominant arm between different subjects does not influence the understanding of LSF. In other words, the mean velocity of the signer does not carry information about linguistic meanings of the movements.
The other result is the difference in the mean velocity of movement of the dominant arm between lexical signs, SASS, and transitions as shown in the histogram of Figure 5. The mean of the mean velocity of the four subjects in the lexical signs, SASS, and the transitions are respectively 0.72 m/s, 0.81 m/s and 0.91 m/s with standard deviation of 0.2 m/s, 0.2 m/s and 0.2 m/s. Thus, the mean velocity in the lexical signs is lower than in the SASS, which is lower than in the transitions.

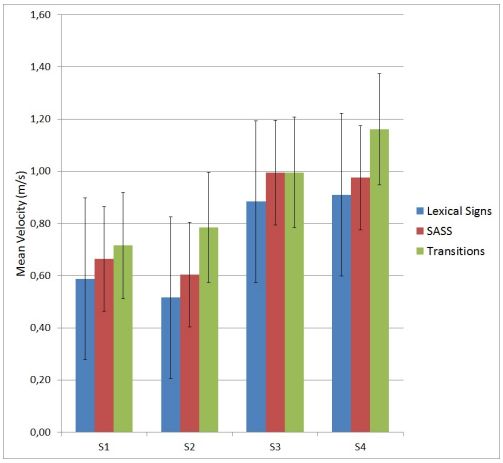

Figure 5: This histogram shows the mean velocity of the four subjects for lexical signs, SASS and transitions

Concerning the transition velocity, an explanation could be that transitions do not convey any message (information) and then need less control, being faster to perform.
Concerning the difference between the lexical signs and the SASS velocity, an hypothesis could be that most of the time, eye gaze is accompanying SASS depicting signs (Braffort, 2016), which is not the case for lexical signs. That could help performing these gestural units in an easier and then faster way.
Of course, these hypotheses should be confirmed by other studies.
This result is also confirmed by the peak of the maximal velocity. The average of the peak of the speed for lexical signs, SASS, and transition respectively is 1.12 m/s 1.25 m/s and 1.36 m/s with standard deviation of 0.26 m/s, 0.32 m/s and 0.26 m/s. This confirms that the velocity in lexical signs and SASS is lower than in transitions.
In conclusion, we can assume that there is more control over arms movements during the signs, being lexical or iconic (SASS) than during transitions.
Therefore, models should be able to produce signs at various paces while preserving the same spatiotemporal organisation. These speeds should also take into account the difference between the types of signs and the transitions, which may be done by means of two parameters that could be tunable in our models in order to change the overall pace of LSF movements.

6.2. Motor control in LSF

Another analysis in progress is to check whether classical laws established in the human motor control literature apply to LSF. That means that we ask the following question: Do classical invariants remain valid during LSF movements? If these laws apply in LSF, one may conceivably assume that classical motor control principles, such as minimum effort or maximum smoothness criteria, may have shaped LSF and must be incorporated into LSF production models. Alternatively, it is possible that LSF requirements led signers to deviate from such classical principles in order to produce very peculiar kinematics of the hands and deliver linguistic meaning. Ongoing investigations will attempt to answer such questions, which is made possible thanks to the creation of a corpus of 3D data of LSF.
One other current focus is related to the law of up-down asymmetries, which states that point-to-point upward movements decelerate for a longer time compared to downward movements, in particular due to the integration of gravity in the motor command driving the limb's motion (Papaxanthis et al., 1998; Gaveau and Papaxanthis, 2011).

6.3. Detection of the dominant hand

An application of using 3D data is the automatic detection of the dominant hand in LSF.
This can be achieved based on the computation of the distances covered by the two hands. By comparing these distances, we can automatically detect the strong hand, which is more active. This computation could be used to automatically feed the metadata related to the signers in annotation software.
We have also studied the variability of this difference across the subjects, by calculating the ratio r between the two distances.

$$r = \frac{D_{weakhand}}{D_{stronghand}} \quad (1)$$

Histogram 2 shows that the ratio (r) is quite stable across the subjects. The global average is r = 0.768 with standard deviation of 0.014.

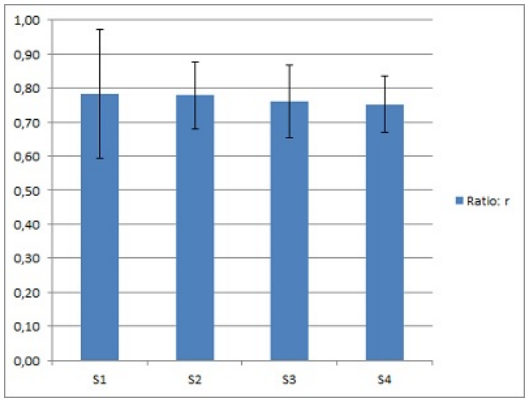

Figure 6: Histogram showing the ratio between the distance covered by the strong hand and the distance covered by the weak hand

7. Conclusion

This paper described the different stages of the constitution of APlus, the first available 3D corpus of LSF, which will be usable in several disciplines. The potential power of analyses based upon the 3D corpus was illustrated. Their main advantage is that they allow to quantify and identify the information-bearing parameters of LSF movements with the aim to use them in the modelling of movements in LSF.

At this moment, the initial part of the corpus, corresponding to the picture description task, has been recorded and fully annotated. The targeted analyses are being completed using the above-mentioned fundamental questions. The first part of the corpus is available on request from authors. The second part has been recorded and annotated.

8. Bibliographical References

Braffort, A., Benchiheub, M.-E.-F., and Berret, B. (2015). APLUS: A 3d corpus of french sign language. In Yeliz Yesilada et al., editors, *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility, ASSETS 2015, Lisbon, Portugal, October 26-28, 2015*, pages 381–382. ACM.

Braffort, A. (2016). Eye gaze in french sign language: a study on depicting signs. In *The 12th Conference on Theoretical Issues in Sign Language Research, 4 - 7 January 2016*, Melbourne, AUSTRALIA.

Burger, B. and Toiviainen, P. (2013). Mocap toolbox - a matlab toolbox for computational analysis of movement data. In *Proceedings of the Sound and Music Computing Conference 2013, SMC 2013, Logos Verlag Berlin*, pages 172–178, Stockholm, Sweden. Logos Verlag Berlin.

Duarte, K. and Gibet, S. (2010). Heterogeneous data sources for signed language analysis and synthesis: The signcom project. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, pages 461–468.

Gaveau, J. and Papaxanthis, C. (2011). The temporal structure of vertical arm movements. *PLoS ONE*, 6(7):e22045.

Héloir, A., Gibet, S., Multon, F., and Courty, N. (2006). Captured motion data processing for real time synthesis of sign language. In Sylvie Gibet, et al., editors, *Proceedings of 6th International Gesture Workshop*, volume 3881 of *Lecture Notes in Computer Science*, pages 168–171. Springer.

Jantunen, T., Burger, B., Weerdt, D. D., Seilola, I., and Wainio, T. (2012). Experiences collecting motion capture data on continuous signing. In Onno Crasborn, et al., editors, *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)*, pages 75–82, Istanbul, Turkey, May. Paris: ELRA.

Lefebvre-Albaret, F. (2010). *Traitement automatique de vidéos en LSF Modélisation et exploitation des contraintes phonologiques du mouvement*. Theses, Université Paul Sabatier - Toulouse III, October.

Lu, P. and Huenerfauth, M. (2010). Collecting a motion-capture corpus of american sign language for data-driven generation research. In *Proceedings of the NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, pages 89–97, Los Angeles, California, June. Association for Computational Linguistics.

Lu, P. and Huenerfauth, M. (2012). Cuny american sign language motion-capture corpus: First release. In Onno Crasborn, et al., editors, *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)*, pages 109–116, Istanbul, Turkey, May. Paris: ELRA.

Malaia, E., Borneman, J., and Wilbur, R. B. (2008). Analysis of asl motion capture data towards identification of verb type. In Bos Johan et al., editors, *Semantics in Text Processing. STEP 2008 Conference Proceedings*, volume 1 of *Research in Computational Semantics*, pages 155–164. College Publications.

Papaxanthis, C., Pozzo, T., and Stapley, P. (1998). Effects of movement direction upon kinematic characteristics of vertical arm pointing movements in man. *Neuroscience Letters 253*, pages 103–106.

Segouat, J. and Braffort, A. (2009). Toward the study of sign language coarticulation: methodology proposal. *Advances in Computer-Human Interactions*, pages 369–374.

Tyrone, M. E., Nam, H., Saltzman, E., Mathur, G., and Goldstein, L. (2010). Prosody and movement in american sign language: A task-dynamics approach. Speech Prosody 2010, 100957, pages 1–4.