Building French Sign Language Motion Capture Corpora for Signing Avatars

Sylvie Gibet
Université Bretagne-Sud, Laboratoire IRISA, France
sylvie.gibet@univ-ubs.fr

Abstract

This paper describes four corpora that have been designed and built in our research team. These corpora have been recorded using motion capture (MoCap) and video equipment, and annotated according to multi-tiers linguistic templates. Each corpus has been designed for a specific linguistic purpose and is dedicated to data-driven synthesis, by (i) replacing signs or groups of signs within an utterance, (ii) replacing phonetic or phonological components and in this way modifying the grammatical or semantic aspects of the phrase, or (iii) altering prosody in the produced sign language utterances.

Keywords: Corpus, MoCap, Sign Language, Signing avatar

1. Introduction

The design of traditional corpora for linguistic analysis aims to provide living representations of sign languages across deaf communities and linguistic researchers. Most of the time, the sign language data is video-recorded and then encoded in a standardized and homogenous structure for open-ended analysis (statistical or phonological studies). With such structures, sign language corpora are described and annotated into linguistic components, including phonology, morphology, and syntactic components (Johnston and de Beuzeville, 2009; Crasborn and Zwitserlood, 2008; Efthimiou and Fotinea, 2007; Wolfe et al., 2011; Hanke et al., 2012).

Conversely, motion capture (MoCap) corpora provide researchers the data necessary to carry on finer-grained studies on movement, thus allowing precise, and quantitative analysis of sign language gestures as well as sign language (SL) generation. One the one hand, motion data serves to validate and enforce existing theories on the phonologies of sign languages. By aligning temporally motion trajectories and labelled linguistic information, it thus becomes possible to study the influence of the movement articulation on the linguistic aspects of the SL, including hand configuration, hand movement, coarticulation or synchronization within intra and inter phonological channels. On the other hand, generation pertains to sign production using animated virtual characters, usually called signing avatars.

Although MoCap technology presents exciting future directions for SL studies, tightly interlinking language components and signals, it still requires high technical skills for recording, post-processing data, and there are many unresolved challenges, with the need to simultaneously record body, hand motion, facial expressions, and gaze direction.
Therefore, there are still few MoCap corpora that have been developed in the field of sign language studies. Some of them are dedicated to the analysis of articulation and prosody aspects of sign languages, whereas recent interest in avatar technology has led to develop corpora associated to data-driven synthesis. In particular, (Lu and Huenerfauth, 2014) collected an ASL corpus and discussed how linguistic challenges in ASL generation could be addressed through this corpus. To improve avatar movement, kinematic and linguistic cues were retrieved from motion capture data and incorporated into a data-driven technique, thus leading to a more realistic animation (Mcdonald et al., 2016). More recently, a MoCap dataset on French sign language (LSF) has been collected (Limsi and CIAMS, 2017). 25 pictures are described in a spontaneous way, and first analysis are conducted. However these corpora do not capture simultaneously the multiple channels involved in SL gestures, i.e. body movement, hand configuration, facial expression, and gaze direction, which remains an important challenge and is necessary to address these highly coded sign languages using multi-tiers linguistic elements, both for recognition (Dilsizian et al., 2014) or synthesis (Gibet et al., 2011).

In this article, we describe four motion capture corpora in French sign language that were designed and built in our research team during the last decade. The technical aspects of the MoCap acquisition are described. Then the linguistic rules that guide the corpora design for the synthesis of new sentences in LSF are discussed and illustrated with examples. Most of these linguistic issues are applied to data-driven generation. However, in all our studies, it is important to emphasize that we adopted a synthesis-by-analysis approach. That is to say, the improvement of our synthesis models led us progressively to refine our methods of segmentation and labeling, to better understand the mechanisms responsible for the formation of the signs, as well as the processes of coarticulation (Naert et al., 2017).

2. Motion Capture Datases

Four corpora in French Sign Language (LSF) and their corresponding MoCap databases have been designed by a team of researchers that includes linguists and computer scientists, Hearing and Deaf. Before describing these corpora we describe hereafter the motion capture databases that were collected over the last ten years in the context of national research projects, with different objectives concerning the linguistic aims and the level of avatar synthesis.

Projects	Markers Capture device	Cameras number	frequency Hz	Databases	Year	Size min
HuGEx	24 (body) 2 Cybergloves 39 (face)	12	120	TRAIN METEO	2005	10’ 40’
SignCom	43 (body) 2x6 (hands) 41 (face)	12	100	SignCom	2009	60’
Sign3D	40 (body) 2x19 (hands) 40 (face) gaze direction	16	100	Sign3D	2013	10’

Table 1: MoCap databases built at IRISA.

2.1. Mocap Devices and Experimental protocols

Different MoCap setups and experimental protocoles were defined in the context of three projects. For all of them, we used a Vicon MX infrared camera technology to capture the 3D displacements of a set of markers. The main differences between the setups of the projects are the number of cameras, of markers, and the frequency of acquisition. Table 1. gives an overview of the projects with the experimental setups. For capturing precisely movements of the hands and facial expressions, it is necessary to use more cameras and to place them closer to the subject so that they can detect all the markers. This is why we increased the number of cameras as we gained experience and mastered the motion capture systems. In addition, the frequency of acquisition has to be large enough to capture the subtle variations of the movements, for example when changing a facial expression, or moving rapidly one hand. We therefore tried to determine the appropriate frequency of acquisition, trying to keep a good compromise between the spatial accuracy and the speed of the cameras. Finally, in all our experimental setups, we considered pairing MoCap with video recordings, assuming that parallel recordings would aid in the ulterior data annotation processes.

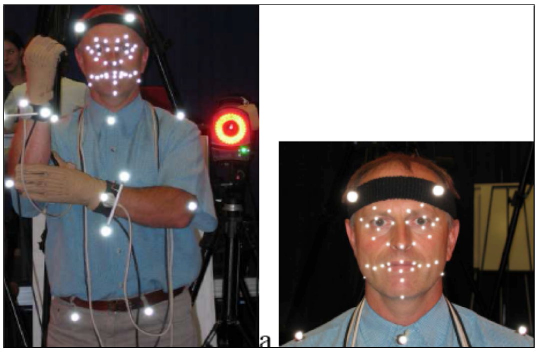

Figure 1: Photo of the MoCap settings in *HuGEx* project.

2.1.1. *HuGEx* project

In *HuGEx* project (Gibet et al., 2006), the Vicon system was composed of 12 infrared cameras cadenced at 120Hz. For the body movements 24 reflective markers were placed on standardized anatomical landmarks. We also recorded facial expressions using 39 small semi-spherical markers (3mm) at locations compliant with Mpeg4 specification. As we had no experience with hand capture, hand movements were recorded using two Cyber gloves (Ascension technologies), each one composed of 22 sensors (see Fig. 1). The fusion of the different signals (body, left and right hand) was realized after reconstruction and synchronization (resampling at 60 Hz). The different information sources (body and hands) were then converted into BVH format. During the recording session, about forty minutes of LSF gestures were captured on one expert deaf signer. This one, who was a trainer in LSF, signed on texts that he himself transcribed into a sequence of glosses. Two databases were built: (i) the *TRAIN* database (about 10 min), aimed at building sentences with predefined replaceable parts; (ii) the *METEO* database (about 40 min), aimed at studying the variation in prosody of the LSF phrases. For both datasets, the mean duration of a sequence was about 60 seconds.

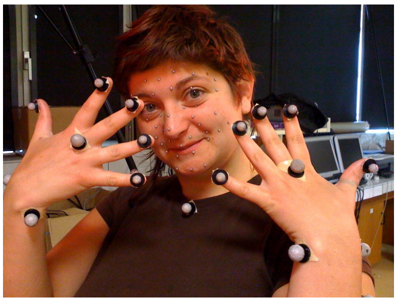

Figure 2: Photo of the MoCap settings in *SignCom* project.

2.1.2. *SignCom* project

The *SignCom* project (Gibet et al., 2011) also used a Vicon MX system with 12 high definition cameras to capture the movements of our LSF signers at a frequency rate of 100Hz. We had 43 markers for the body, and 41 markers of small diameter for the face. Instead of using data gloves which lack precision and exhibit significant drift when used for a long time, we captured the hand movements by fixing 6 markers per hand (see Figure 2).

As for the previous projects, we used an additional video camera to have video recordings in addition to MoCap

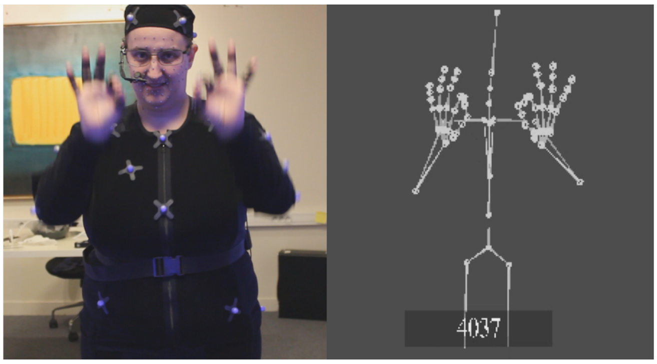

Figure 3: Photo of the MoCap settings in *Sign3D* project.

data. This is necessary for annotation. Body movements, hand and finger movements, and facial expressions were recorded simultaneously by the Vicon system. Two professional deaf linguists signing in LSF designed the corpus, and learned it by heart. During the recording session, the information was presented to the deaf signers through images projections, so that the signers were able to recall the scenarii without reading any text-based translation. 68 motion sequences were recorded on one signer. This constitutes the *SignCom* database containing about one hour of MoCap data. From this data (in C3D format), a skeleton was reconstructed, and the data (body and hand movements, as well as facial expressions) was converted into the formats BVH and FBX.

2.1.3. *Sign3D* project

The *Sign3D* project (Gibet et al., 2015) used a Vicon T160 system with 16 high definition cameras at a frequency rate of 100 Hz, combined with a head-mounted oculometer (MocapLab MLab 50-W), designed to track the gaze direction. Facial expressions, body and finger motions were again simultaneously recorded. For recording precisely hand movements and hand configurations, we used a much larger number of markers (19 per hand against 6 in the *SignCom* project). The LSF gestures of one expert deaf signer were recorded for about 10 minutes to form the dabase *Sign3D*. The motion capture settings associated to the skeleton reconstruction is illustrated in Figure 3. The motion skeleton data including body and hand movements) was converted into the FBX format. The facial expressions data was converted into blendshape coefficients.

3. Corpora Design

We describe hereafter the various corpora that we have designed in the context of the former research projects. Through these projects, we developed a complete concatenative data-driven synthesis pipeline that enables the assembling of motion elements, from signs and parts of sentences, to motion chunks retrieved from different channels and body parts (hand movements, hand configurations, body movements, facial expressions, and gaze direction), representing phonetic or phonological components.

Our corpora follow the objectives of synthesis by replacing signs or groups of signs in sentences, by composing phonetic or phonological components, and finally by analyzing and generating prosody in sentences carried out in different stylistic contexts.

3.1. Motivation

For the purpose of corpus design, three main questions have been addressed in the three former projects. The first one concerns the corpus content itself and the compromise that exists between breadth and depth in its design. The second question concerns the nature of the sign variability which is of paramount importance if we want to create new sentences in different discourse contexts. The third question concerns the acted or spontaneous nature of the produced SL utterances.

Concerning the first question, we wanted to have control over the signs or gloses that appeared into the corpora, and therefore we would prefer a limited vocabulary of given signs, and multiple instances for each sign played in different discourse contexts. We also chose to incorporate standard signs into our lexicon, as they were easier to handle for synthesis. Given the difficulty of capturing large corpora (a tedious and time-consuming process, both in terms of capture, post-processing, and annotations), we also opted for a limited set of utterances or sequences of signs. Therefore, in parallel with the design of our sentences, we had to think deeply about the mechanisms of editing signs and constructing new sentences.

The question of variability can be approached in different ways: (i) by constructing sentences containing the same signs appearing in different contexts; (ii) by repeating the sentences several times and with different subjects; and (iii) by enriching the initial corpus with new constructed sentences.

To answer the third question, in all our corpora, the scenarii were scripted by deaf persons, and the produced sign language utterances were acted. Table 2 gives an overview of the corpora, indicating the level of annotation, the topic, and the linguistic application.

3.2. Replacing Signs or Groups of Signs in Sentences

The first experimental ideas for synthesizing new statements from original sign language data were to insert replaceable parts into a sentence, such as signs, or groups of signs. This was first achieved in the *HuGEx* project where the corpus was composed of a set of phrases expressing incident reports relatively to the railway traffic, with a set of additional signs representing French towns. Two excerpts are shown below. The brackets delimit the variable elements.

> The train from [CITY] to [CITY] is delayed by [DURATION], due to [CAUSE].
> The train [NUMBER] is being prepared; the starting lane will be displayed in [DURATION].

It was then possible to build programmable sentences by choosing the departure and arrival cities ([CITY]) among a given set of pre-recorded cities, the number of the train ([NUMBER]), or the nature of the incident ([CAUSE]) belonging to the following set: a technical incident / bad

Databases	Annotation, Segmentation	Nature	Linguistic Application
TRAIN	gloss	Train incident Towns, numbers	Fill-gap-synthesis
Sign3D	phonetic, phonology, gloss	Urban services Places, schedules, event rates	Phonological synthesis Hand movement analysis
SignCom	phonetic, phonology, gloss	Recipes Interactive dialogs	Pattern-based synthesis Coarticulation analysis
METEO	gloss	Weather forecast Emotional variations	Prosody analysis

Table 2: Corpora in LSF.

weather / personal accident).

In the *Sign3D* project, we collected a corpus of French sign language utterances, describing various events (exhibitions, inaugurations, cultural events) taking place in various buildings and monuments (swimming pool, theater, town hall, museum, etc.), indicating the opening and closing hours, entrance fees, their location relative to each other, and the potential occurrence of an incident (weather, work, etc.). In this latter corpus, the aim was also to build new sentences by replacing signs (hours, buildings, etc.), or groups of signs (events, incident causes). To preserve the linguistic coherence of the LSF statements, while optimizing the number of variants of the different sentences, the corpus was designed by declining a limited set of syntactic patterns (the brackets delimit the variable elements).

> The [LOCATION] is [ABSOLUTE OR RELATIVE LOCATION]; it opens at [TIME] and closes at [TIME].
> Access is [PAYING / FREE], [ENTRY CONDITION].
> The [EVENT] in [LOCATION] is moved to [LOCATION] due to [CAUSE].
> In case of [CAUSE], the [EVENT] in [LOCATION] will be moved to [LOCATION].

where the variables ([LOCATION], [TIME], etc.) may be replaced by values belonging to a given set of signs. Thirteen sample sentences were then signed by a deaf LSF expert. This corpus can easily be extended by enriching it with the synthesized variant sentences. This represents about 10 minutes of continuous LSF.

3.3. Altering Phonological Components of Signs

The objective of the *SignCom* project was also to design new utterances in LSF, and to animate a virtual signer, using both raw motion and annotated data. The idea was similar to the previous projects, but instead of manipulating signs, the aim was to re-assemble phonetic or phonological elements of signs, while keeping the global coherence and realism of the produced sequences. The corpus contains three thematic scenarii: the *Cocktail* monologue, and the *Galette* and *Salad* dialogues. The scenarii were scripted by two expert deaf people who designed the scenes using comic stories that were displayed on the back wall of the room, thus avoiding lowering the head for reading the scenarii. Both deaf people trained for several days before the recording sessions, hence they executed the motion as acted sequences. A total of 68 sequences was captured and annotated, following a multi-tier template with different levels of labeling (gloss, phonological and phonetic elements), for each hand separately, and for the two-hands.

As signed languages are by nature spatial languages, forming sign sequences requires a signer to understand a set of spatial-temporal grammatical rules and inflection processes. These processes have oriented the range of LSF signs recorded for the project. This brought us to include a number of various linguistic inflection mechanisms into the corpus that allow for creating novel sentences from our original corpus. After defining a delimited vocabulary, we chose to introduce spatial references (for example depicting and indicating verbs which are modulated in the context of dialog situations), changes in hand configurations, and changes in hand movements.

3.3.1. Spatial references: directional verbs and pointing movements

We included in the dataset directional verbs and depicting verbs as well as personal and possessive pronouns. This gave us the possibility to build new sentences by conjugating the verbs. For example, the sign INVITE can be modified grammatically to become ”I invite you”, ”You invite me”, etc. Our vocabulary thus contains several instances of the directional verbs shown in Table 3. A certain number of pointing gestures in different parts of the signing space were also included. These targets are labeled with their 3D location.

3.3.2. Changes in hand configurations

Many hand configurations, possibly associated to verbs, allow for designing different objects, or indicate *size* or

Salad	Cocktail	Directional verbs
22 × SALAD	8 × COCKTAIL	GIVE
20 × PRO-1	8 × DRINK	TAKE
19 × WHAT	7 × GLASS	PROPOSE
8 × PLATE	7 × FRUIT	INVITE
6 × TOMATO	3 × ORANGE	COMMUNICATE
12 × POUR	3 × JUICE	PUT
11 × WANT	7 × FILL	EXPLAIN
9 × CHEVRE	7 × THERE-IS	QUESTION
9 × VARIOUS	3 × NEXT	
3 × AVOCADO	4 × ALCOHOL	
5 × ADD	2 × WITHOUT	

Table 3: Two first columns: some tokens with their occurrence in the *Salad* and *Cocktail* scenarii (*SignCom* corpus); Third column: directional verbs mainly found in the dataset.

shape specifiers. Given our inclusion of signs that take multiple hanshapes, like GIVE, we introduced in the corpus different hand configurations from other signs that can be substituted to the original handshapes. In the case of GIVE, most often signed in our dataset as if the signer was handing a glass to someone, a hanshape substitution could yield addition meanings, such as giving a piece of paper or giving an object with a cylindrical shape. In particular, the expression GIVE A GLASS is performed in our corpus with glasses of different sizes and forms (for example a large glass, a thin long glass, or a champagne flute).

3.3.3. Changes in movement kinematics

Analyzing hand movements has shown regular shapes (bell-shapes) which differ whether they belong to strokes (within-sign) or transitions (inter-signs). Moreover, for strokes, toward-target movements differ from backward movements (Duarte and Gibet, 2010). These observations have led us to introduce many kinematic variations of movements in the corpus, so that it becomes possible to analyze and annotate these patterns, and to retrieve the appropriate movement from the database that preserve the temporal coherency of the reconstructed phrase.
The corpus also contains reversal verbs, as for example the sign GIVE which can be reversed in the sign TAKE, or the sign LIKE which can be reversed in DO-NOT-LIKE in LSF.

3.3.4. Composition process to build new sentences

An overview of the most frequent tokens in the Cocktail and Salad scenarii is provided in Table 3. With this variety and frequency of our related lexemes, we are able to produce a number of novel utterances based on the thematic subjects. Examples of construction of new sentences from the above transformations is shown in the following examples:

> I GIVE-YOU a THIN-GLASS (1)
> I TAKE a LARGE-GLASS (2)
>
> I LIKE FRUIT JUICE (3)
> I DO-NOT-LIKE ORANGE JUICE (4)

In this first example, only the right arm is involved. The movement (2) begins at the position where the movement (1) ends; the direction of movement is reversed, and the shape of the hand is changed to handle a big glass instead of a thin glass. In the second example, different channels are combined, by keeping the torso/lower-body/left-arm of one sequence (3), and substituting the head, facial expression and right arm movements of another sequence (4). The sign DO-NOT-LIKE is reversed from the sign LIKE. In such a composition process, the spatial constraints should be preserved, in particular the sign ORANGE should be executed near the corresponding body part (head), whatever the torso or the head orientation is. This clearly reveals that the combination process should be driven at a more abstract level, expressed by rules or constraints incorporated into the animation engine.

3.4. Altering the Prosody of Sentences

The objective of the *HuGEx* project was to animate with a data-driven approach a virtual signer endowed with expressive sign language gestures. Our attention focused on the prosody of the LSF gestures, and on its influence on the semantic comprehension. The corpus *METEO* was composed of a set of sentences describing weather forecasts, performed with different variations of expressiveness: *neutral*, *angry*, *emphasis*, and *tired*. The mean duration of a sequence was 60 seconds. We took as referent performance the first sequence performed according to neutral style. An example retrieved from the corpus is given below.

> Today, July 6th, here is the weather forecast. In the morning, clouds will cross Brittany. In the afternoon, it will rain. Tomorrow, the sun will shine. It will be hot and dry.

An analysis of the LSF prosody was achieved on the expressive sentences, through a temporal alignment process using an adaptive dynamic time warping algorithm (Héloir et al., 2006). Using machine learning techniques, it would be possible to learn the sequences performed with different styles and then transfer the style of one sequence into another one.

4. Conclusion

In this article, we described four corpora with different linguistic purposes that have been designed and built in our research team over the last ten years. These corpora were recorded using MoCap data, post-processed and manually annotated, so that they could be used for different goals: linguistic analysis, automatic annotation, or generation.

With the increasing interest of linguistic or computer science researchers using sign language motion capture data, there is a need to provide motion capture databases that can be shared by the different communities. Following the approach adopted by other research teams in movement sciences that have made available raw motion, videos, and tools, with exchangeable data formats, we want to share our experience on the design of corpora and the construction of MoCap databases. We also propose to make available soon our LSF MoCap corpora.

Concerning the annotated data, we have used schemes inspired from the linguistic community, and we are currently enriching these schemes by developing automatic annotation methods. These annotated schemes (manual or automatic) with the documentation explaining the structure and the coding system of the annotation should also be shared by the different research communities.

Other more focused corpora are also currently designed and collected in our research team. They are dedicated to the automatic annotation of two specific channels: facial expressions and hand configurations, and will be used for animating a signing avatar (Naert et al., 2018).

5. Acknowledgements

The motion capture datasets presented in this paper have been realized, thanks to the *HuGEx* project funded by the CNRS, the *SignCom* project funded by the French National Research Agency (ANR), and the *Sign3D* project, funded by the French Ministry of Industry.

6. References

Crasborn, O. and Zwitserlood, I. (2008). Annotation of the video data in the Corpus NGT. Technical report, Department of Linguistics and Center for Language Studies, Radboud University, Nijmegen, the Netherlands, November.

Dilsizian, M., Yanovich, P., Wang, S., Neidle, C., and Metaxas, D. (2014). A new framework for sign language recognition based on 3d handshape identification and linguistic modeling. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Duarte, K. and Gibet, S. (2010). Reading between the signs: How are transitions built in signed languages? In *Theretical Issues in Sign Language Research (TILSR 2010), Indiana, USA*.

Efthimiou, E. and Fotinea, S.-E. (2007). GSLC: Creation and annotation of a Greek Sign Language corpus for HCI. In *Universal Access in Human Computer Interaction. Coping with Diversity*, volume 4554 of *Lecture Notes in Computer Science*, pages 657–666. Springer, Berlin.

Gibet, S., Héloir, A., Courty, N., Kamp, J., Gorce, P., Rezzoug, N., Multon, F., and Pelachaud, C. (2006). Virtual agent for deaf signing gestures. In *AMSE, Journal of the Association for the Advancement of Modelling and Simulation Techniques in Enterprises (Special edition HANDICAP)*, pages 127–136.

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. In *Transactions on Interactive Intelligent Systems*, volume 1. ACM.

Gibet, S., Lefebvre-Albaret, F., Hamon, L., Brun, R., and Turki, A. (2015). Interactive editing in French sign language dedicated to virtual signers: requirements and challenges. *Universal Access in the Information Society*, 15(4):525–539.

Hanke, T., Matthes, S., Regen, A., and Worseck, S. (2012). Where does a sign start and end? Segmentation of continuous signing. *Language Resources and Evaluation Conference*.

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

Johnston, T. and de Beuzeville, L. (2009). Researching the linguistic use of space in Auslan: Guidelines for annotators using the Auslan corpus. Technical report, Department of Linguistics, Macquarie University, Sydney, June.

Limsi and CIAMS. (2017). Mocap1. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Lu, P. and Huenerfauth, M. (2014). Collecting and evaluating the cuny asl corpus for research on american sign language animation. *Comput. Speech Lang.*, 28(3):812–831, May.

Mcdonald, J., Wolfe, R., Wilbur, R., Moncrief, R., Malaia, E., Fujimoto, S., Baowidan, S., and Stec, J. (2016). A new tool to facilitate prosodic analysis of motion capture data and a data- driven technique for the improvement of avatar motion. In *Proceedings of the 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining; Language Resources and Evaluation Conference*, volume 7.

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 *8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community; Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japon, May 12.

Wolfe, R., McDonald, J., Schnepp, J., and Toro, J. (2011). Synthetic and acquired corpora: Meeting at the annotation. In *Workshop on Building Sign Language Corpora in North America, Washington, DC*.