Which Picture?
A Methodology for the Evaluation of Sign Language Animation Understandability

Vonjiniaina Domohina MALALA[1], Elise PRIGENT[1], Annelies BRAFFORT[1], Bastien BERRET[2]

[1]Limsi, CNRS, Paris Saclay University – [2]CIAMS, Paris-Sud University, Paris Saclay University

[1]LIMSI, Campus d'orsay bat 508, F-91405 Orsay cedex, France

[2]CIAMS, Campus d'orsay bat 335, F-91405 Orsay cedex, France

{elise.prigent, annelies.braffort}@limsi.fr, bastien.berret@u-psud.fr

Abstract

The goal of our study is to explore which information is essential to understand virtual signing. To that aim, we developed an online test to assess the comprehensibility of four different versions of signers: a baseline version with a real human signer, a most complete version of a virtual signer, and two degraded versions of a virtual signer (one with non-visible hands and one without movements of head/trunk). Each video showed the description of a picture in French Sign Language (LSF). After having seen the video, participants had to find which picture had been described among 9 pictures displayed. The originality of our approach was to include two types of confusable pictures on the response board. One was supposed to induce errors by confounding the lexical signs and the other by confounding the spatial structure of the picture. In this way, we explored the effect of hiding hands and blocking trunk/head on the comprehension of lexicon and spatial structure.

Keywords: Sign Language, SL animation, Virtual signer, Evaluation, Visual Perception

1. Introduction

This paper deals with the evaluation of Sign Language (SL) generation technology. We focused on French SL (LSF) generation based on the animation of a Virtual Signer (VS) using real human movements captured from a motion capture system (Benchiheub et al., 2016).

SL is a visuo-gestural language that uses eyes, face, torso, arms and hands movements to convey meaning. With actual technology, all these movements cannot be replicated accurately and faithfully in virtual signing whereas they are likely necessary for understanding the signing content. This may account for the poor understandability of virtual signers by Deaf people. Therefore, to produce understandable signing, we must determine which visual information is most critical and has to be perfectly animated on the VS. Thus, the questions raised in this study are: What motion information is essential to understand virtual signing? To what extent does the manipulation, simplification or withdrawal of some information affect understanding?

To provide objective answers to these questions, we have to design a method allowing to acquire quantitative measures about visual perception and comprehension of VS (with different qualities) by participants.

In this paper, we introduce the method that was designed and used in a cognitive psychology study related to the visual perception of movements in SL.

The paper is structured as follows: section 2 is dedicated to a review of studies in psychology and computer science about perception and comprehension of human and virtual signers; section 3 details our methodology and the design of the platform used; section 4 proposes a discussion and gives an example of the kind of interesting results that have been acquired thanks to this method.

2. Visual perception of Sign Language

2.1 Perception of Human Signers

Emmorey, Thompson & Colvin (2009) have shown that both Deaf native signers (Deaf people with SL as native language) and hearing beginning signers (who had completed between 9 and 15 months of SL instruction) look at the observed signer's face more than 80% of the time. But these authors also showed that beginning signers move more frequently their attention from the face to the hands than native signers. Native signers would focus their attention on the eyes while retaining the ability to integrate the information from the manual parameters with peripheral vision (Morford et al., 2008). Despite native Deaf signers focus their attention on the face, they recognize more quickly the signs conveyed by the hands than beginning signers (Morford & Carlson, 2011). This confirms that this is rather the peripheral vision of signers that is used to perceive the rapid movements of the hands and fingers, while central vision is used to perceive movements located on the face.

Thus, according to Muir (2005), a good spatial resolution of the image at the face level (with good temporal resolution maintained throughout the video) is necessary for understanding SL videos. For this author, it may be possible to reduce the quality of the peripheral region, including the body and hands (when away from the face), while retaining the quality of the perceived video.

2.2. Perception of Virtual Signers

The use of Virtual Signers (VS) brings many advantages over videos of real signers. They are anonymous and can be interactive (Kipp et al., 2011b). Nevertheless, their usability is limited by the low level of comprehension by the observers (Kennaway et al., 2007). Most VS are developed by researchers who are not experts of SL linguistics and who tend to create "pleasant" VS, sometimes forgetting that the VS is also a language that convey informa-

tion with movements of the articulators that can be very precise. That is why we must identify which component of the model needs more precision, for optimal understanding (Kipp et al., 2011a).
Moreover, the creation of VS animations remains a difficult task because the movements of the different parts of the body must be well synchronized and it is difficult to reproduce all the spatiotemporal parameters of SL, in particular the non-manual parameters. Criticisms regarding VS have often pointed out these parameters (gaze, facial expression, movement of the mouth, head and bust) (Kipp et al., 2011a). Paradoxically, great attention is usually put on the movements of the hands to facilitate the understanding of SL. For instance, Alexanderson & Beskow (2015) proposed to use a low-cost technology using fewer markers in the animation of the movements of the hands of VS, thereby obtaining a recording of hand movements of less complete/accurate. The results of this study showed that despite the reduction in the amount of information, the comprehensibility and the clarity of the signs was not altered compared to the animation with more markers.
Regarding eye movements, an eye-tracking study showed that when native Deaf people observe human signers, the fixation time of the face is greater than when they observe VS. Accordingly, there is less gaze displacements between the face and the body when observing a human signer rather than a VS (Kacorri et al., 2014).
These previous studies suggest two main results: first, there is a difference about visual exploration (of face and body) of users when observing human or virtual signers, with different levels of quality; second, the comprehension but also the visual exploration used can differ as function of the observer's SL expertise.
In order to determine the parameters of SL that must be modelled more precisely for the optimisation of the VS, we explored observers' comprehension of different types of signers: human, virtual with different qualities by manipulating different relevant parameters. We also explored the impact of the observer's SL expertise on their comprehension. Our VS was animated using motion capture of a human signer and not using synthetic animations, thus guaranteeing data very close to the initial human signing.

3. Methodology

From previous studies, we know that user-based evaluation of SL generation comprehensibility requires many precautions during the design step, regarding the identification of the socio-linguistic profile of the participants and avoid using to much text in order to keep the participants concentrated on SL.
There is no standard process for assessing the comprehensibility of an LSF statement. Generally, simple categories are proposed to evaluate globally the understandability and naturalness, sometimes grammatical correctness, using for example numeral scales or glosses[1] as possible responses given by the participants (Kipp et al., 2011a).
Huenerfauth et al. (2008) have proposed an original process, that consists to use short movies. Each movie gives a dynamic interpretation of an utterance such as "The man walk next to the woman". The participant had to match each SL animation with one movie among three. This approach can provide a more reliable rating of understandability, but it cannot be used for any kind of utterances.

3.1 Our set-up: an online test with complete and degraded animations

As previously mentioned, our objective is to determine which parameters of SL must be modelled more precisely for the optimisation of the VS. To evaluate quantitatively the relevance of different body parts on the SL comprehensibility, a method consists to alter the animations and compare the perception, such as in (Huenerfauth & Lu, 2010) regarding the location of signs, or in (Gibet et al., 2011) for facial expression and gaze. We have used the same kind of method, while trying to add a more reliable way to measure the understandability.
We used Cuxac's model (2000) to determine the relevant parameters. According to this model, we hypothesized that the lack of handshapes should result in more difficulty to identify the lexical signs, while the lack of body and head movements should result in more difficulty to figure out the global structure of the picture, which is described by "showing", in the signing space in front of the signer, the spatial organisation of the picture scene, implying in many cases rotations and movements of the head and the torso. Because many studies focused on lexical signs comprehension, we propose here to explore the comprehension of signs related to more depicting structures, such as size and shape descriptions or localisation of entities in the signing space. Hence, we measured the impact of two degradations of the virtual signers on the comprehension of LS description, and more precisely on the comprehension of lexical signs and depicting signs[2] respectively. In our study, one version of the animation was realised by hiding the hands and the other one by blocking the trunk and head movements on all degrees of freedom.

In order to allow a relatively large number of persons to participate and collect enough data to conduct statistical analyses, we created an online test via a LimeSurvey server (a web application that enables users to develop and publish online surveys, collect responses and export the resulting data).
We asked participants to watch 8 videos containing picture descriptions in LSF (see section 3.1.). After each video, the participant had to choose the picture described among a set of 9 pictures (see, section 3.2.). This online test was send to Deaf Signer, Hearing Signer and Hearing Non Signer using mailing lists or social networks in France.

[1] Word or set of word expressing the same concept (or at least the closest), i.e. the gloss SCIENCE for the lexical sign representing the concept of *science.*

[2] These types of signs are often referred to as 'classifier' signs. See (Liddell, 2003) for a detailed definition of depicting sign.

3.2 Stimuli

For the creation of visual stimuli, we used the LSF corpus called MOCAP1 (Benchiheub et al., 2015), which in particular contains videos and motion capture data of the description of pictures. An expert made an annotation of the corpus, segmenting the gestural units in the videos, then identifying the lexical signs and the depicting structures, especially those showing size and shape of objects, localisation of objects or spatial relations between objects. Based on these annotations, we chose the 4 descriptions with approximately the same number of lexical and depicting signs in order to create our stimuli (these 4 pictures are illustrated in Figure 1). In this corpus, in addition to a camera, 3D recordings of the movements were made using a motion capture system (Optitrack). These recordings were then used to animate the VS.

The physical appearance of the VS, the color of his skin, the clothes and the background of the video were chosen to have as much resemblance as possible to the original video. From the 3D recordings, a Deaf computer graphist created 3 different versions of VS (Figure 2):

- A complete animation without modification (Complete VS, Figure 2.b).
- A degraded animation with hands hidden by spheres (Handless VS, Figure 2.c)
- A degraded animation by freezing the trunk and head movements (Blocked VS, Figure 2.d)

Since the human signer had no markers on the fingers, the computer graphist manually animated fingers and facial expressions by using the rotoscopy method. So these 3 versions of VS presented facial expressions based on those displayed by the real signer. Thus, for each of the 4 pictures (Figure 1), we obtained 4 videos of the description corresponding to 4 types of signers: human signer, complete VS, handless VS, and blocked VS (Figure 2).

3.3 Modality of response

After each video, a response board with 9 pictures was displayed, and the participant had to choose the picture described in the video. For each video description, 8 confusable pictures were carefully chosen according to the expert's annotations mentioned previously. 4 pictures presented similarities in the lexicon (related to the objects present in the scene), 4 others in the global structure of the described picture.

More precisely, the 9 pictures were composed of:

- 1 picture corresponding to the correct response, the one described in the video (Figure 3: n°6),
- 4 confusable pictures with similar spatial structure (Figure 3: n° 4, 5, 7, 8),
- 4 confusable pictures with similar lexical elements (Figure 3: n° 1, 2, 3, 9).

The same response board was displayed in the two conditions "real signer" and "virtual signer", but with a different ordering of the pictures in the response board.

Because we displayed two degraded versions of VS, we could test whether the handless VS induces more confusion for pictures with similar lexical elements and, conversely, whether the blocked VS induces more confusion for pictures with similar structures.

Figure 1. Example of pictures used to elicite descriptions in the MOCAP1 LSF corpus. These are the 4 ones used in our study.

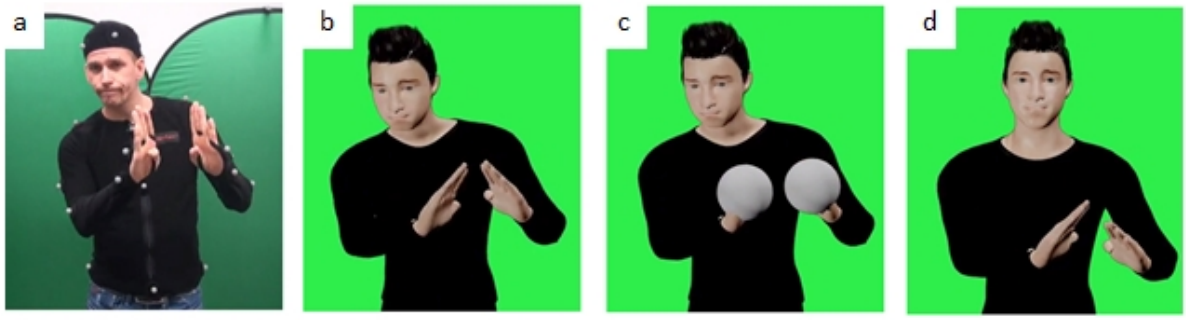

Figure 2. Extracts of videos in the 4 conditions: a) Real signer; b Complete VS; c) Handless VS; d) Blocked VS.

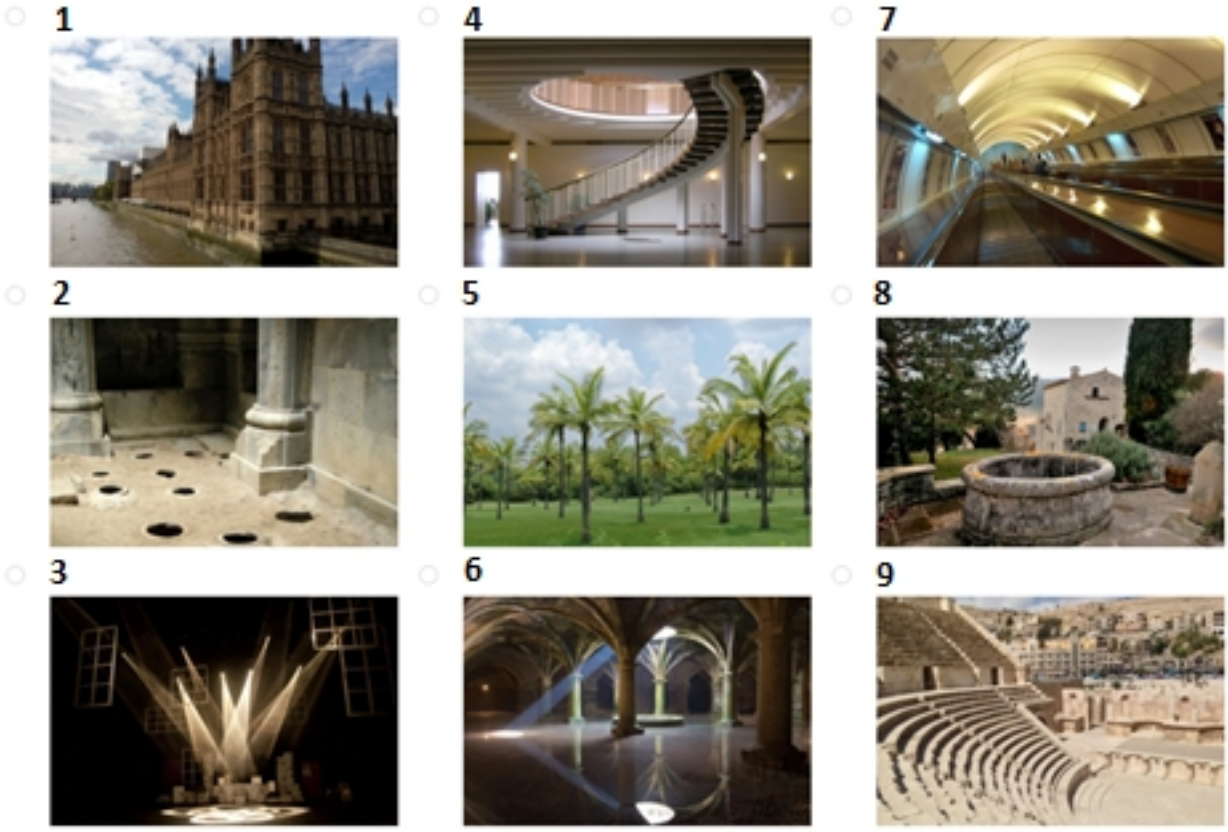

Figure 3. Example of a response board. The picture 6 is the right answer. Pictures 1, 2, 3, 9 are supposed to induce lexical errors. Picture 1 could induce the lexical sign “water”; 2: “hole”; 3: “light”; 9: “Roman”. Pictures 4, 5, 7, 8 are supposed to induce structural errors. Picture 4 could induce a description that illustrates the shape of a hole in the ceiling; 5: shape of pillars and ceiling; 7: shape of vaulted ceiling; 8: shape of the well edge.

3.4 Procedure

We asked participants to run the test on desktop computer or laptop to ensure good viewing conditions of the videos, because the screens of mobile phones or touch pads are too small.

The test lasted about 10 minutes. The first page of the test provided instructions and explanations on the process, and the informed consent of the participants. The test was performed anonymously. Instructions were presented in written form and with a video translation in LSF to facilitate accessibility and understanding of the task by Deaf persons who might present reading difficulties.

Prior to the comprehension test, participants were asked to indicate their age, gender, nationality, hearing status and expertise in LSF. That constituted 3 groups: Deaf Signer (DS), Hearing Signer (HS), Hearing Non Signer (HNS). There were no Deaf Non Signer participants. Depending on the group to which they belong, the participants were directed toward different questions. For example, a DS or HS participant had to answer questions related to his/her level in LSF (according to the European Common European Framework of Reference for Language). A DS

participant had to answer questions related to their place and age of learning of LSF, etc.

The comprehension test was composed of two blocks, the first with 4 videos description of VS and the second with 4 videos description of human signer. Within each block, the order of the 4 videos was randomized. We created 3 different versions of the VS (complete, handless and blocked). So, each participant was randomly oriented to one of the 3 versions of VS (complete, handless or blocked) in block 1 and all participants watched the same block 2 of human signers. Each video lasted between 20 and 25 seconds. For each trial, one video of description of picture in LSF was displayed. Participants could only view the video once. The "play", "stop" and "progress bar" commands were deactivated and backtracking on the web page was not allowed. To prepare the participant to the video, a 4-seconds countdown was displayed before his beginning. Once the video is finished, the participant clicked on the “next” button to access the page containing the response board with 9 pictures (1 good response, 4 “structural” confusable pictures and 4 “lexical” confusable pictures). The participant had to click on the picture that he thinks correspond to the description in the previous video, and then click on the "next" button to move on to the next video. Before starting the comprehension test, a familiarization trial, not included in statistical analyses, was displayed, composed of a video of the same person describing a different image than the ones used in the test.

At the end of the test, the participants had the opportunity to get the number of correct answers they got and to give their impressions on the test and on the VS by using a text field.

4. Discussion

The results of this specific study is not the focus of this paper. But to say a word, the very first analysis gives some insights about the role of movement in understandability. For example, even with the degraded versions of the VS animations and even in the group of hearing non signers participants, some of them were able to find the good response. On the other hand, even Deaf and hearing signers could be disturbed by the degraded versions of the VS but not necessarily in the same manner.

The discussion here is more on the design of the test. The originality of our method is to propose a link between the information degraded in the stimuli (here, hands hidden and blocking trunk and head movements) and the confusable pictures displayed in the response board. This design allows us to measure the impact of the degradation of a visual information on the comprehension of the message, and more precisely on the comprehension of two types of signs, lexical and depicting, by analysing errors made by the participants. The results may provide interesting conclusions both for linguistic and computer science domains. First, they could serve linguistic models by providing information about the relative importance of the movement of specific body parts (face, hand and bust) for the various type of sign (lexical or depicting). Second, this study may provide some new guidelines for the animation of VS. Because synthetic animation of VS does not allow to accurately replicate all the movements of a human signer, a simplification is necessary. So, this kind of study can propose recommendations about simplification of one motion parameter rather than another as a function of the message produced (e.g. lexical signs or depicting signs) and of the expertise in SL of the participants.

Moreover, we have yet some inputs on the way the test could be improved. Actually, near 200 participants have completed the online test. We had much more participants, but responses from non-French participants and those who did not perform the test until the end were excluded from the analyses. Thus, the design of an online test allows to get a sufficient number of participants as well as to perform robust and reliable statistical analysis. However, we have not a balanced size of participants in the 3 groups (Deaf signers, Hearing signers and Hearing non signers). There were less Deaf participants. It also appeared that several participants had only a smartphone and thus were rejected from the test for which we asked to use a desktop computer or a laptop. Also, a limitation is that the participants had the opportunity to give their impressions on the test only by text. That could be a brake for Deaf people who present writing difficulties. We plan to add the possibility to post impressions via a video in follow-up studies of this type.

Another difficulty is related to the duration of the descriptions. They lasted between 20 and 25 seconds, which may seem short, but they contain an important number of elements (between 17 and 26 depicting and lexical signs). Overall, the descriptions are already quite complex. Therefore, there may be some memorisation issues that are part of the difficulty of the task. We assume that this effect, which is the same for all the participants, has no influence on the results and interpretations. However, this is perhaps one of the reasons why some participants did not complete the test. It would be interesting in the future to think about a more playful way of presenting the test, like a serious game for example.

5. References

Alexanderson, S., & Beskow, J. (2015). Towards Fully Automated Motion Capture of Signs – Development and Evaluation of a Key Word Signing Avatar. *ACM Transactions on Accessible Computing*, 7(2), 1-17.

Benchiheub, M.-E., Berret, B., & Braffort, A. Collecting and Analysing a Motion-Capture Corpus of French Sign Language, 7th International Conference on Language Resources and Evaluation - Workshop on the Representation and Processing of Sign Languages (LREC-WRPSL 2016), Portoroz, Slovenia, 7-12.

Cuxac, C. (2000). La Langue des Signes Française (LSF) : les voies de l'iconicité, Paris-Gap, Ophrys, Bibliothèque de *Faits de Langues* n°15 -16, 2000

Emmorey, K., Thompson, R., & Colvin, R. (2009). Eye Gaze During Comprehension of American Sign Language by Native and Beginning Signers. *Journal of Deaf Studies and Deaf Education*, 14(2), 237-243.

Gibet, S., Courty, N., Duarte, K. & Naour, T. (2011), The SignCom System for Data-driven Animation of

Interactive Virtual Signers: Methodology and Evaluation, *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1:1, 6:1--6:23.

Huenerfauth, M., Zhao, L., Gu, E., Allbeck, J.: Evaluating american sign language generation by native ASL signers. *ACM Transactions on Access Computing* 1(1), 1–27 (May 2008)

Huenerfauth, M. & and Lu, P. (2010), Modeling and synthesizing spatially inflected verbs for American sign language animations, Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility (ASSETS'10), Orlando, Florida, USA, 99-106.

Kacorri, H., Harper, A., & Huenerfauth, M. (2014). Measuring the perception of facial expressions in american sign language animations with eye tracking. In International Conference on Universal Access in Human-Computer Interaction (p. 553–563). Springer.

Kennaway, J. R., Glauert, J. R. W., & Zwitserlood, I. (2007). Providing signed content on the Internet by synthesized animation. *ACM Transactions on Computer-Human Interaction*, 14(3).

Kipp, M., Heloir, A., & Nguyen, Q. (2011a). Sign language avatars: Animation and comprehensibility. In International Workshop on Intelligent Virtual Agents (p. 113–126). Springer.

Kipp, M., Nguyen, Q., Heloir, A., & Matthes, S. (2011b). Assessing the deaf user perspective on sign language avatars, The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility, 107-114.

Liddell, S. K. (2003). Grammar, Gesture, and Meaning in American Sign Language. Cambridge: Cambridge University Press.

Morford, J. P., & Carlson, M. L. (2011). Sign Perception and Recognition in Non-Native Signers of ASL. *Language Learning and Development*, 7(2), 149-168

Morford, J. P., Grieve-Smith, A. B., MacFarlane, J., Staley, J., & Waters, G. (2008). Effects of language experience on the perception of American Sign Language. Cognition, 109(1), 41-53.

Muir, L. J. (2005). Perception of Sign Language and Its Application to Visual Communications for Deaf People. *Journal of Deaf Studies and Deaf Education*, 10(4), 390-401.

Alexanderson, S., & Beskow, J. (2015). Towards Fully Automated Motion Capture of Signs – Development and Evaluation of a Key Word Signing Avatar. *ACM Transactions on Accessible Computing*, 7(2), 1-17.

Benchiheub, M.-E., Berret, B., & Braffort, A. Collecting and Analysing a Motion-Capture Corpus of French Sign Language, 7th International Conference on Language Resources and Evaluation - Workshop on the Representation and Processing of Sign Languages (LREC-WRPSL 2016), Portoroz, Slovenia, 7-12.

Cuxac, C. (2000). La Langue des Signes Française (LSF) : les voies de l'iconicité, Paris-Gap, Ophrys, Bibliothèque de *Faits de Langues* n°15 -16, 2000

Emmorey, K., Thompson, R., & Colvin, R. (2009). Eye Gaze During Comprehension of American Sign Language by Native and Beginning Signers. *Journal of Deaf Studies and Deaf Education*, 14(2), 237-243.

Gibet, S., Courty, N., Duarte, K. & Naour, T. (2011), The SignCom System for Data-driven Animation of Interactive Virtual Signers: Methodology and Evaluation, *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1:1, 6:1--6:23.

Huenerfauth, M., Zhao, L., Gu, E., Allbeck, J.: Evaluating american sign language generation by native ASL signers. *ACM Transactions on Access Computing* 1(1), 1–27 (May 2008)

Huenerfauth, M. & and Lu, P. (2010), Modeling and synthesizing spatially inflected verbs for American sign language animations, Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility (ASSETS'10), Orlando, Florida, USA, 99-106.

Kacorri, H., Harper, A., & Huenerfauth, M. (2014). Measuring the perception of facial expressions in american sign language animations with eye tracking. In International Conference on Universal Access in Human-Computer Interaction (p. 553–563). Springer.

Kennaway, J. R., Glauert, J. R. W., & Zwitserlood, I. (2007). Providing signed content on the Internet by synthesized animation. *ACM Transactions on Computer-Human Interaction*, 14(3).

Kipp, M., Heloir, A., & Nguyen, Q. (2011a). Sign language avatars: Animation and comprehensibility. In International Workshop on Intelligent Virtual Agents (p. 113–126). Springer.

Kipp, M., Nguyen, Q., Heloir, A., & Matthes, S. (2011b). Assessing the deaf user perspective on sign language avatars, The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility, 107-114.

Liddell, S. K. (2003). Grammar, Gesture, and Meaning in American Sign Language. Cambridge: Cambridge University Press.

Morford, J. P., & Carlson, M. L. (2011). Sign Perception and Recognition in Non-Native Signers of ASL. *Language Learning and Development*, 7(2), 149-168

Morford, J. P., Grieve-Smith, A. B., MacFarlane, J., Staley, J., & Waters, G. (2008). Effects of language experience on the perception of American Sign Language. Cognition, 109(1), 41-53.

Muir, L. J. (2005). Perception of Sign Language and Its Application to Visual Communications for Deaf People. *Journal of Deaf Studies and Deaf Education*, 10(4), 390-401.

6. Language Resource References

MOCAP1 corpus (2015), distributed via Ortolang. perennial identifier https://hdl.handle.net/11403/mocap1