Best Practices for Conducting Evaluations of Sign Language Animation

Matt Huenerfauth

Rochester Institute of Technology, Golisano College of Computing and Information Sciences

matt.huenerfauth@rit.edu

Hernisa Kacorri

The Graduate Center, CUNY, Doctoral Program in Computer Science

hkacorri@gc.cuny.edu

Abstract

Automatic synthesis of linguistically accurate and natural-looking American Sign Language (ASL) animations would make it easier to add ASL content to websites and media, thereby increasing information accessibility for many people who are deaf. Based on several years of studies, we identify best practices for conducting experimental evaluations of sign language animations with feedback from deaf and hard-of-hearing users. First, we describe our techniques for identifying and screening participants, and for controlling the experimental environment. Finally, we discuss rigorous methodological research on how experiment design affects study outcomes when evaluating sign language animations. Our discussion focuses on stimuli design, effect of using videos as an upper baseline, using videos for presenting comprehension questions, and eye-tracking as an alternative to recording question-responses.

Keywords

Deaf and Hard of Hearing; Emerging Assistive Technologies; Research & Development

Introduction

Standardized testing in the U.S. has revealed that many deaf adults have lower levels of English reading literacy (Traxler). If the reading level of the text is too complex on websites, closed-captioning, or other media, these adults may not understand the content. More than 500,000 people in the U.S. use American Sign Language (ASL) as a primary means of communication (Mitchell et al. 328-329), and worldwide, nearly 70 million people use a sign language (World Federation). So many individuals can benefit from information conveyed in this form; traditionally, this is done by displaying videos of human signers. However, automatically synthesized animations have advantages, including enabling frequent updates without re-recording a human performer and supporting dynamic content generation.

Researchers working on this technology must evaluate whether animations are grammatically correct and understandable, typically through participation of signers, e.g. (Gibet et al. 18-23; Kipp et al. 107-114; Schnepp et al. 250). Until recently, the field has lacked rigorous methodological research on how experiment design affects study outcomes. Our lab has conducted several research projects, surveyed in (Huenerfauth, Learning), to investigate experimental methodologies. Informed by this prior work, including hundred of hours of studies with deaf participants, this article summarizes best practices for conducting such evaluations.

Discussion

Identifying and Screening Participants

When humans evaluate a language generation system, it is important for them to be native speakers of that language: proper screening is needed to ensure that these judges are sufficiently critical of the system’s output (Neidle 15). An ideal participant is a “native signer,” someone who learned ASL in early childhood through interactions at home or through significant time in a school environment using ASL. We have effectively advertised for such participants in metropolitan areas through distributing messages to online groups and email lists and through hiring recruiters from the local Deaf community. We have also found that it is ineffective to screen potential participants by asking questions such as "How well do you sign?," "Are you a native signer?," or "Is ASL your first language?" (Huenerfauth et al. 213-214). Such questions could be misinterpreted as asking whether the individual feels personally oriented toward Deaf culture. We have instead found it effective to ask whether the potential participant has had life experiences typical of a native signer: "Did you grow up using ASL as a child?," "Did your parents use ASL at home?," "Did you attend a residential school where you used ASL?," etc.

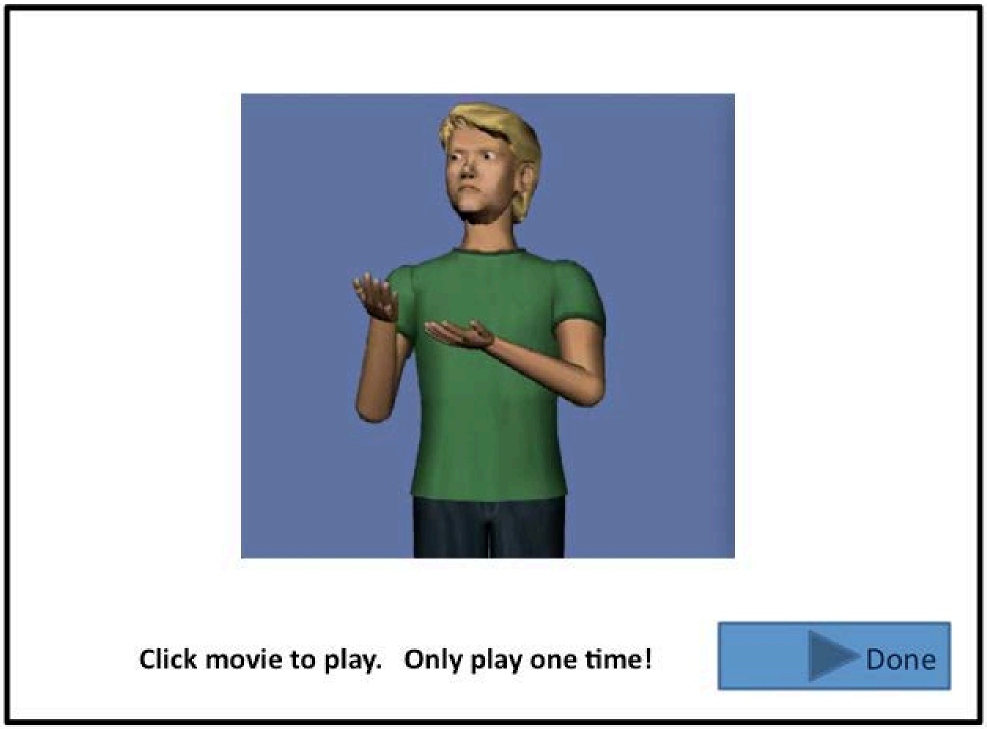
Controlling the Experimental Environment

When seeking grammaticality judgments from signers, it is important to minimize environmental characteristics which may prompt signers to code-switch to more spoken-language-like forms of signing or accept such signing as grammatically correct (Neidle 15). Many signers are accustomed to switching to such signing in interactions with hearing individuals. To avoid this, participants should be exposed only to fluent sign language during the study (Huenerfauth et al. 213-214). Instructions should be signed by another native signer. If possible, participants should be immersed in a sign language environment prior to the study, e.g., engaging in conversation in fluent ASL prior to the study. If interpreters are required, they should possess near-native sign language fluency (Huenerfauth et al. 213-214).

As with any study, users must feel comfortable criticizing the system being evaluated. In this context, it is important that the participant not feel that anyone responsible for the system is sitting with them while they critique it – or else they may not feel as comfortable offering negative opinions about the system. If a native signer is "hosting" the study, it is helpful for this person to present themselves as an "outsider" to the technical team that had created the animations being evaluated (Huenerfauth et al. 216).

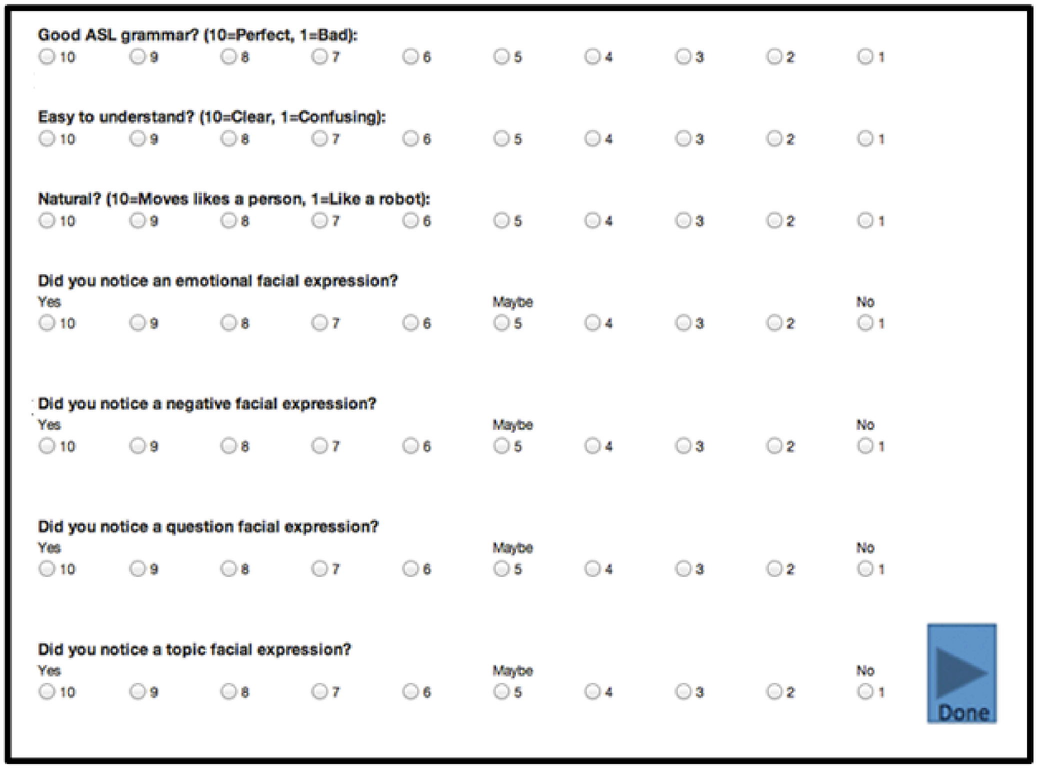
Engineering Stimuli for Studies

Inventing stimuli that contain specific linguistic phenomena and measure whether participants understand the intended information is challenging – but necessary for effectively evaluating ASL animations. For instance, in prior work, we have described how to engineer animation stimuli that can be interpreted (ambiguously) in different ways, depending on whether a particular aspect of the sentence was successfully understood by the participant (e.g., whether a particular ASL facial expression was correctly perceived). In this way, comprehension questions can be invented that specifically measure whether this aspect of the animation was correct, thereby enabling an evaluation of that specific issue (Kacorri, Lu, and Huenerfauth, Evaluating 514-516). To aid researchers, we have published our methods for designing stimuli for a variety of linguistic phenomena in ASL, and we have also released a collection of stimuli for evaluating ASL facial expressions (Huenerfauth and Kacorri).



Click movie to play. Only play one time!

Fig.1a. Example screens of a stimulus.



Good ASL grammar? (10=Perfect, 1=Bad):

○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Easy to understand? (10=Clear, 1=Confusing):

○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Natural? (10=Moves likes a person, 1=Like a robot):

○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Did you notice an emotional facial expression?

Yes Maybe No

○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Did you notice a negative facial expression?

Yes Maybe No

○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Did you notice a question facial expression?

Yes Maybe No

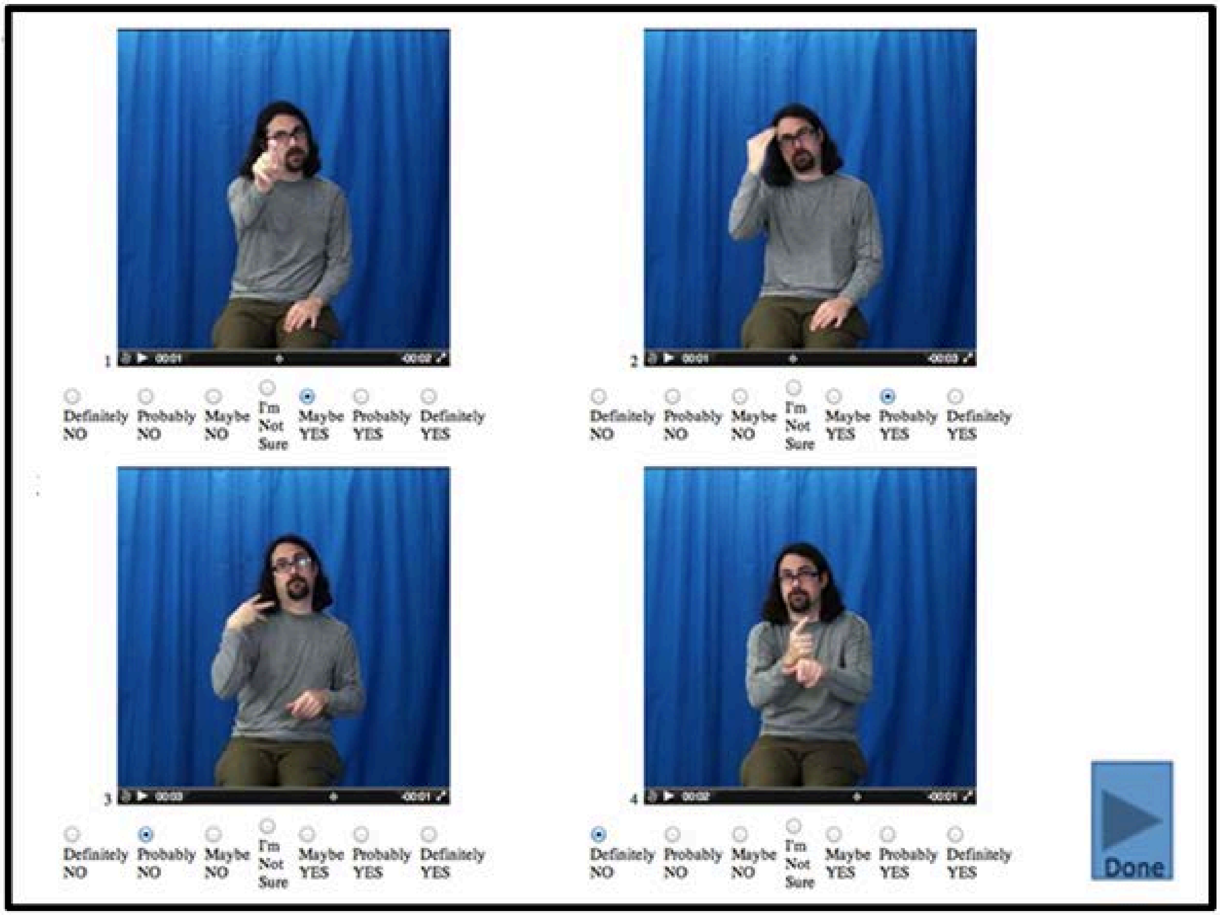
○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Did you notice a topic facial expression?

Yes Maybe No

○ 10 ○ 9 ○ 8 ○ 7 ○ 6 ○ 5 ○ 4 ○ 3 ○ 2 ○ 1

Fig.1b. Example screens of subjective evaluation questions.



1
○ Definitely NO ○ Probably NO ○ Maybe NO ○ I'm Not Sure ● Maybe YES ○ Probably YES ○ Definitely YES

2
○ Definitely NO ○ Probably NO ○ Maybe NO ○ I'm Not Sure ○ Maybe YES ● Probably YES ○ Definitely YES

3
○ Definitely NO ● Probably NO ○ Maybe NO ○ I'm Not Sure ○ Maybe YES ○ Probably YES ○ Definitely YES

4
● Definitely NO ○ Probably NO ○ Maybe NO ○ I'm Not Sure ○ Maybe YES ○ Probably YES ○ Definitely YES

Fig.1c. Example screens of comprehension questions in a user study.

Engineering Subjective Evaluation Questions for Studies

To measure user's satisfaction with sign language animations after viewing stimuli, we have asked participants to answer subjective questions concerning grammatical correctness, ease of understanding, and naturalness of movement of the virtual human character (Huenerfauth and Lu 174-176). To ensure that they are clearly communicated, these questions are explained in sign language, and participants select answer choices on Likert scales. In (Huenerfauth et al. 216-217), we observed that the scores measuring grammaticality, naturalness, and understandability were moderately correlated, which was understandable since the grammaticality and naturalness of an animation could affect its perceived understandability. In other studies, we have also asked participants to rate on Likert scales how confident they were that they had noticed specific phenomena of interest in the animations, e.g., a specific facial expression. A questionnaire with both types of Likert-scale questions and their answer choices was released in (Huenerfauth and Kacorri).

Engineering Comprehension Evaluation Questions for Studies

While it is relatively easy to ask a participant to rate subjectively whether they believe a particular animation stimulus was understandable, we have observed low correlation between a user's subjective impression of the understandability of a sign language animation and his/her actual success at answering comprehension questions about that animation (Huenerfauth et al. 216-217). It is for this reason that we have made efforts to include an actual comprehension task (either a comprehension question about information content in the stimulus or a matching task that the user must perform based on this information). We have discussed how users' perceived understandability scores are not an adequate substitute for this actual comprehension data.

To obtain reliable scores, researchers must ensure that spoken-language skills are not necessary for participants to understand comprehension questions or answer choices. In prior work, we have presented comprehension questions in sign language, e.g. using videos of a native singer or high quality animations created by a native signer. We found that presenting questions as video or high-quality animation did not affect comprehension scores (Kacorri, Lu, and Huenerfauth, Effect 22-27). To present answer choices, we have successfully used image matching (Huenerfauth et al. 216-217), clip-art illustrations for answer choices (Huenerfauth, Evaluation 132-133), or definitely-no-to-definitely-yes scalar responses (Kacorri, Lu, and Huenerfauth, Effect 2-27; Lu and Kacorri 187-188).

As discussed above, for comprehension scores to be meaningful, they must be engineered to probe whether participants have understood the intended information specifically conveyed by the aspect of the animation that the researcher wishes to evaluate. This is particular challenging for non-manual components of animation, e.g. facial expressions (Kacorri, Lu, and Huenerfauth, Evaluating 514-516). To aid other researchers in conducting studies, we have released to the research community a set of 192 comprehension questions for ASL stimuli with facial expressions (Huenerfauth and Kacorri).

Use of Baselines for Comparison

In general, the *absolute* scores recorded from questions in a study are difficult to interpret unless they can be considered *relative* to some baselines for comparison. This is because the absolute scores in a study may depend on a variety of factors beyond the animation-quality, e.g., the difficulty of the stimuli and the comprehension questions, participants' memory skills, etc. Thus, in addition to the to-be-evaluated version of an animation stimulus, we include other stimuli in a study so that the relative scores can be compared.

As a "lower baseline" for comparison, we have found it effective to present users with a version of the ASL animation that differs from the stimuli by excluding only the features being evaluated, e.g., if we are evaluating a method to add a particular facial expression to an animation, the lower baseline will lack this facial expression (illustrated in Fig. 2). A good "upper baseline" should represent an "ideal" system output and may consist of a high-quality computer animation or a video recording of a human signer (performing identical sentences to the virtual human in the animations). We compare both approaches in (Lu and Kacorri 183-189; Kacorri, Lu, and Huenerfauth, Effect 2-22).

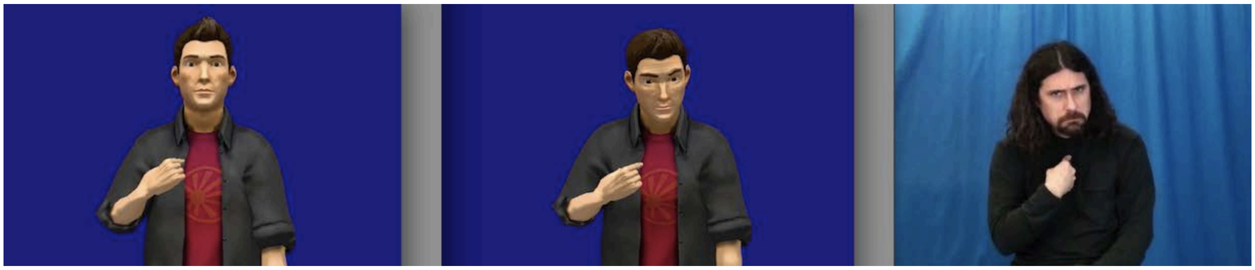


Fig. 2. Example of three types of stimuli in a user study: i) animation without facial expressions as lower baseline, ii) animation with facial expressions to be evaluated, and iii) video of human signer as upper baseline.

Eye-tracking metrics in Evaluation Studies

Researchers sometimes need to measure users' reactions to animations without obtrusively directing participants' attention to the new features being incorporated; in such cases, we have investigates the use of eye-tracking technologies to evaluate stimuli (Kacorri, Harper, and Huenerfauth, Comparing; Kacorri, Harper, and Huenerfauth, Measuring 549-559). We divided the screen region where the stimuli appear to three areas of interest: "Upper Face", "Lower Face", and "Hands". Figure 3 illustrates these areas of interest for the animations of the virtual character and for the videos of the human signer in our experiment. We found that the time-normalized fixation trail length metric should be utilized if seeking an eye metric that correlates with participants' subjective judgments about ASL videos or animations.

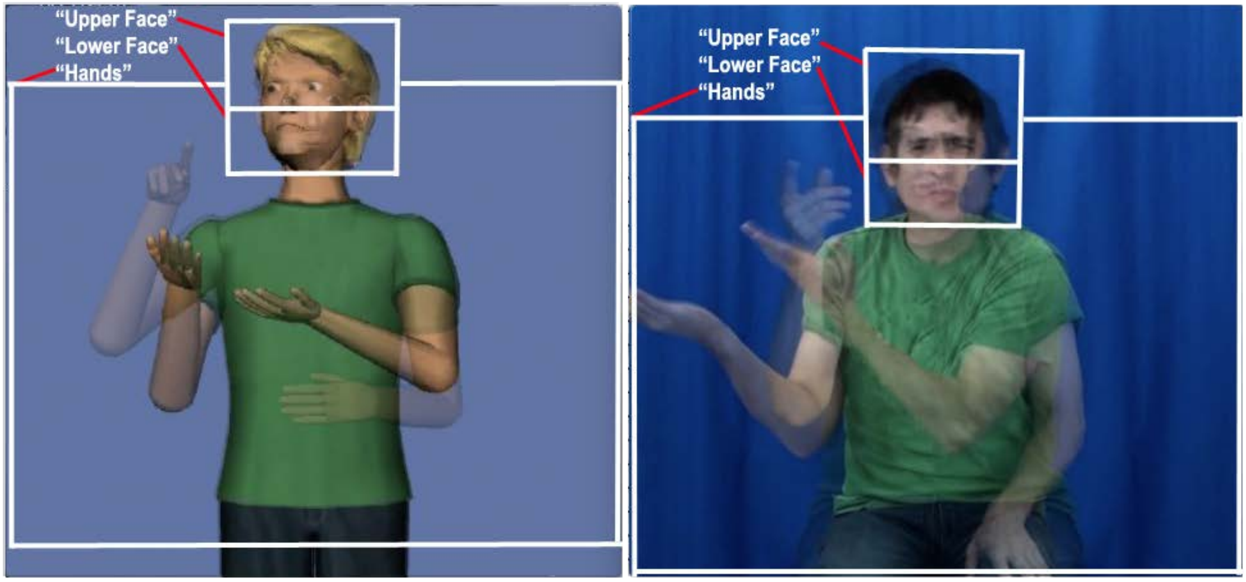


Fig. 3. Screen regions for the upper face, lower face, and hands AOIs.

Conclusions

The article is designed to serve as a resource for future researchers who must design experimental evaluations of technology with deaf participants, and it offers guidance on how to effectively evaluate sign language animation technologies.

Works Cited

Gibet, Sylvie, Nicolas Courty, Kyle Duarte, and Thibaut Le Naour. "The SignCom System for Data-driven Animation of Interactive Virtual Signers: Methodology and Evaluation." ACM Transactions on Interactive Intelligent Systems (TiiS) 1.1 (2011): 6.

Huenerfauth, Matt. "Evaluation of a Psycholinguistically Motivated Timing Model for Animations of American Sign Language." Proceedigns of the 10th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2008). New York: ACM Press, 2008.

Huenerfauth, Matt. "Learning to Generate Understandable Animations of American Sign Language." 2nd Annual Effective Access Technology Conference. 2014.

Huenerfauth, Matt. “A Linguistically Motivated Model for Speed and Pausing in Animations of American Sign Language.” ACM Transactions on Accessible Computing 2.2 (2009): 9.

Huenerfauth, Matt, and Pengfei Lu. 2012. “Effect of Spatial Reference and Verb Infection on the Usability of American Sign Language Animations.” Universal Access in the Information Society 11.2 (2012): 169-84.

Huenerfauth, Matt, and Hernisa Kacorri. “Release of Experimental Stimuli and Questions for Evaluating Facial Expressions in Animations of American Sign Language.” Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel, The 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland. 2014.

Huenerfauth, Matt, Liming Zhou, Erdan Gu, and Jan Allbeck. “Evaluating American Sign Language Generation Through the Participation of Native ASL Signers.” Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2007). New York: ACM Press, 2007.

Kacorri, Hernisa, Allen Harper, and Matt Huenerfauth. “Measuring the Perception of Facial Expressions in American Sign Language Animations with Eye Tracking.“ Universal Access in Human-Computer Interaction. Lecture Notes in Computer Science 8516 (2014): 549-59. Switzerland: Springer International Publishing, 2014.

Kacorri, Hernisa, Allen Harper, and Matt Huenerfauth. “Comparing Native Signers Perception of American Sign Language Animations and Videos via Eye Tracking.” Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2013). New York: ACM Press, 2013.

Kacorri, Hernisa, Pengfei Lu, and Matt Huenerfauth. “Effect of Displaying Human Videos During an Evaluation Study of American Sign Language Animation.” ACM Transactions on Accessible Computing 5.2 (2013): 4.

Kacorri, Hernisa, Pengfei Lu, and Matt Huenerfauth. “Evaluating Facial Expressions in American Sign Language Animations for Accessible Online Information.” Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques for eInclusion, Lecture Notes in Computer Science 8009 (2013): 510-19. Switzerland: Springer International Publishing, 2013.

Kipp, Michael, Quan Nguyen, Alexis Heloir, and Silke Matthes. "Assessing the Deaf User Perspective on Sign Language Avatars." Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2011). New York: ACM Press, 2011.

Lu, Pengfei, and Hernisa Kacorri. “Effect Of Presenting Video As A Baseline During An American Sign Language Animation User Study.” Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2012). New York: ACM Press, 2012.

Mitchell, Ross E., Travas A. Young, Bellamie Bachleda, and Michael A. Karchmer. “How Many People Use ASL in the United States? Why Estimates Need Updating.” Sign Language Studies 6.3 (2006): 306-35. Gallaudet University Press. 2006.

Neidle, Carol Jan, ed. The syntax of American Sign Language: Functional categories and hierarchical structure. MIT Press, 2000.

Schnepp, Jerry C., Rosalee J. Wolfe, John C. McDonald, and Jorge A. Toro. "Combining emotion and facial nonmanual signals in synthesized american sign language." Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2012). New York: ACM Press, 2013.

Traxler, Carol Bloomquist. "The Stanford Achievement Test: National Norming and Performance Standards for Deaf and Hard-of-hearing Students." Journal of Deaf Studies and Deaf Education 5.4 (2000): 337-348. Oxford University Press. 2000.

World Federation of the Deaf. “Sign Language.” 5 July 2014. <http://wfdeaf.org/human-rights/crpd/sign-language>.