Signing Avatar Performance Evaluation within the EASIER Project

Athanasia–Lida Dimou[1]ⓘ, Vassilis Papavassiliou[1], John McDonald [2]ⓘ, Theodoros Goulas[1]ⓘ, Kyriaki Vasilaki[1]ⓘ, Anna Vacalopoulou[1]ⓘ, Stavroula-Evita Fotinea[1]ⓘ, Eleni Efthimiou[1]ⓘ, Rosalee Wolfe [1]ⓘ

[1] Institute for Language and Speech Processing (ILSP)/ ATHENA R.C.

[2] School of Computing, DePaul University

[1]Artemidos 6 & Epidavrou, 15125 Maroussi, Greece
[2] 243 S. Wabash Ave, Chicago, IL 60604, USA
{ndimou, vpapa, tgoulas, kvasilaki, avacalop, evita, eleni_e, rosalee.wolfe,}@athenarc.gr,
jmcdonald@cs.depaul.edu

Abstract

The direct involvement of deaf users in the development and evaluation of signing avatars is imperative to achieve legibility and raise trust among synthetic signing technology consumers. A paradigm of constructive cooperation between researchers and the deaf community is the EASIER project[1], where user driven design and technology development have already started producing results. One major goal of the project is the direct involvement of sign language (SL) users at every stage of development of the project's signing avatar. As developers wished to consider every parameter of SL articulation including affect and prosody in developing the EASIER SL representation engine, it was necessary to develop a steady communication channel with a wide public of SL users who may act as evaluators and can provide guidance throughout research steps, both during the project's end-user evaluation cycles and beyond. To this end, we have developed a questionnaire-based methodology, which enables researchers to reach signers of different SL communities on-line and collect their guidance and preferences on all aspects of SL avatar animation that are under study. In this paper, we report on the methodology behind the application of the EASIER evaluation framework for end-user guidance in signing avatar development as it is planned to address signers of four SLs -Greek Sign Language (GSL), French Sign Language (LSF), German Sign Language (DGS) and Swiss German Sign Language (DSGS)- during the first project evaluation cycle. We also briefly report on some interesting findings from the pilot implementation of the questionnaire with content from the Greek Sign Language (GSL).

Keywords: signing avatar performance1, on-line questionnaire2, evaluation methodology3, signing avatar rating4, signer involvement5, deaf-friendly interfaces6.

1. Introduction

The use of avatars in signed communication can be implemented in multiple communication contexts permitting a significant degree of freedom in content creation and signer anonymization. Avatars offer the advantage of being flexible to editing changes of the signed content and anonymity of the user. These features enable avatars to serve as agents for various interactive environments and communication platforms. However, currently SL avatars have not yet reached a level of performance that would make them acceptable to their end-users.

To identify how human signers perceive and evaluate the performance of an avatar's synthetic signing, within EASIER project, we have developed a shell environment which incorporates an on-line questionnaire for feedback collection. This allows for easy creation of targeted on-line questionnaires to be addressed to signer groups of different SLs to collect feedback on various aspects of interest regarding research work on synthetic signing technology.

The paper reports on the implementation framework of this user involvement methodology, the goal being the steady improvement of animation regarding legibility and clarity of synthetic signing.

In section 2, we present the on-line questionnaire structure along with the methodological approach adopted to optimize its usability and structural design, aiming to eliminate common and uncommon biases.

Starting from the shell questionnaire design, the goal has been to create an environment which would maintain user-friendly characteristics and respect accessibility requirements of its target audience while guarding against bias. To exemplify application of the adopted methodology, in section 3, we also present results from the questionnaire's first pilot implementation with content from the Greek Sign Language (GSL). Finally, section 4 provides a discussion on our goals and up-to-date experience.

2. The EASIER Questionnaire for Avatar Performance Evaluation

The key performance indicators (KPIs) regarding the EASIER avatar performance are clearly user-centric, identified around perceived naturalness and comprehensibility. To encourage user engagement in the evaluation process, the users themselves participated in the development of the questionnaire format from the state of its design. To further facilitate usability of the questionnaire, comprehensibility is subject to a yes/no response, while naturalness is related to a rating scale from 1 to 5 (various aspects of collecting user feedback with similar focus is also reported in (Kipp et al., 2011) and (Kacorri et al., 2015) among others). It becomes also clear that user involvement from early stages of development

---

[1] https://www.project-easier.eu/

becomes mandatory, if both these qualities are to be judged positively during an official evaluation procedure (EUD, 2018; WFD, 2018 on user attitude). Here we present the overall rationale as well as those specific parameters which led decision-making regarding the design of the shell questionnaire environment that allows creation of targeted on-line questionnaires for the evaluation of the various aspects of avatar performance under development, making use of language material from different SLs.

2.1 Questionnaire Content Design

While designing the architecture of the shell on-line questionnaires we considered various parameters which allow for generation of the overall layout of each specific questionnaire. Among the issues to be tackled are decisions as to how the questionnaire should be best distributed to its audience along with the profile of those it would be addressed to. This is directly connected also with the need to regularly address end-users while proceeding with different stages of technological development (Wolfe et al., 2021). Thus, decisions on questionnaire content led to focused, short lasting questionnaire implementations.

One of our main concerns was to balance between a reasonable questionnaire duration (maximum 20 minutes) that would not cause discomfort or fatigue to the participants, and adequate content to provide clear data on the intended user preferences for which feedback is requested. By setting up a viable, easily updated on-line survey we opted to engage in a steady dialogue with signers' communities with respect to novel enhancements in the signing avatar technology.

Having the possibility to adapt the survey outline according to the evaluation requirements at each stage of avatar development was a decisive factor that weighed on the survey framework design. We needed to provide options for one item viewing at a time or head-to-head alternative performance presentations so that viewers can express their preference, but also provide scorings associated with each performance. To test content presentation settings and design adequacy regarding collection of user opinions in view of the project evaluation procedures, the pilot application of the on-line questionnaire involved two distinct avatars and was composed of two parts that address a set of evaluation questions from different angles. In this setting, the first questionnaire part presented the two avatars on the same screen in a head-to-head manner, while the second part presented one avatar at a time. In this way, we had the opportunity to gather user feedback regarding the entire range of options for content presentation and rating mechanisms incorporated in the shell-questionnaire environment. The customizable aspect of this shell environment allows for tailored, easy, and fast content integration on targeted questionnaires, independent from language specific characteristics or context induced particularities.

Special care was taken so that in the questionnaire pages in which two different versions of signing avatar performance appear, these are presented in similar body and face dimensions and against a similar background, to minimize bias in display settings.

2.2 Questionnaire Usability and Technical Features

A major concern was to provide a survey shell fully adapted to the three-dimensional language modality. Considering that language is the principal tool for human interaction, we ensured that all questionnaire parts and items can be accessible with the use of sign language only. Hence, in every stage of the questionnaire participants are provided with instructions as to what they are expected to do and how they may interact with the questionnaire environment in the following three ways:

i. Via SL videos recorded by an L1 signer of the addressed SL community,

ii. Via written text available to be viewed if selected, in a text box below each instructive video, and

iii. Via screen capture videos demonstrating the requested action by the user in the form of a visual manual.

An introductory video presents the scope of the questionnaire, the identity of the research team and a brief description of the EASIER project.

Questionnaire pages make use of color code to indicate user selections. Color is also used to notify for missing actions which are required to be completed in a given page before the user is allowed to move to the next page. Checking graphical signals are also used to help visualization of user selections (
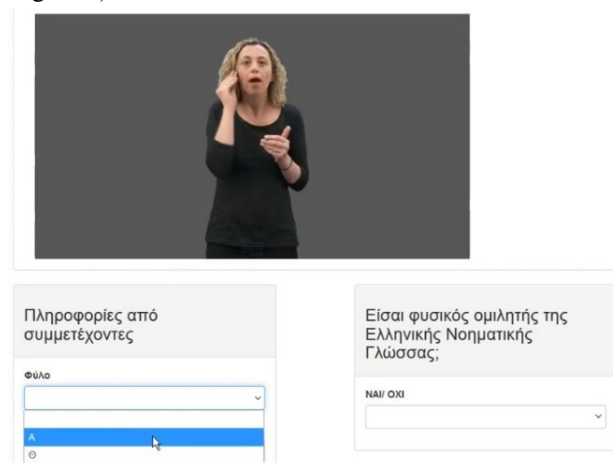
Figure **1**).

Figure 1: Photo from the screen capture video presenting the instruction module with SL video display and visualizing user selections.

Written text instructions to guide user preference selection have also been subject of extensive study aiming to avoid disorienting the users from the focus point of each questionnaire page.

The pilot implementation of the on-line survey was performed with input from the Greek Sign Language (GSL) and was addressed to GSL signers. Therefore, all instructions were linguistically adapted to the target language.

Each on-line questionnaire is available via a URL in which participants can watch avatar productions in the form of embedded videos. Regarding software technologies, the shell questionnaire is created using the open-source Cascading Style Sheets of the Bootstrap Framework. Bootstrap is a framework that allows the creation of responsive, mobile-first web applications. Thus, web applications created by Bootstrap Framework can be executed by most desktop as well as mobile browsers. However, due to the considerable number of images and

videos in the application, participants are encouraged to use Firefox or Chrome for optimum performance. The user interface has been created using HTML5 and JavaScript (jQuery). The database in which participants' answers are stored is MySQL. Php is used to store the data in the database.

3. Aspects of the Pilot Questionnaire Application

For the pilot survey, the questionnaire was divided in two parts, Part A and Part B. Part A targeted GSL user opinion on affect, hand movement, hand, and finger configuration accuracy in isolated signs and in fingerspelling, and Part B targeted smoothness of transition in short phrases. Both parts made use of the EASIER signing avatar "PAULA" (McDonald et al., (2016), (Wolfe et al., 2011) initially developed at DePaul University (http://asl.cs.depaul.edu/), and the Dicta-Sign signing avatar "FRANÇOISE" (Jennings et al., 2010) developed at the University of East Anglia (UEA) (http://vh.cmp.uea.ac.uk).

The linguistic content of the questionnaire was distributed in the following manner:

In Part A the participants were presented with pairs of avatars, head-to-head in randomized order. There were 19 signing instances in all, grouped into 4 categories. For each pair, participants indicated the avatar they preferred and rated the performance quality of both avatars. The four categories were:

(i) Avatar expressivity via inspection of still images of avatar face, in various affect expressions.
(ii) Avatar productions performing signs with varying articulatory formations
(iii) Avatar performance in proper name fingerspelling tasks
(iv) Avatar productions of short phrases composed from previously evaluated isolated signs integrated with signs not yet viewed by participants.

In Part B participants observed one avatar at a time. Each avatar performed a set of signs and short phrases. In this part each of the two avatars displayed different content. The aim here was to lead viewers to focus on specific features of interest in each avatar performance. Tasks included rating each avatar separately in respect to:

(i) Overall hand motion performance,
(ii) Overall body motion performance,
(iii) Head and eyes movement,
(iv) Mouth movement.

The pilot survey has focused on L1 and L2 signers' different preferences of the two avatars. Hence, the sample of the population to which the questionnaire was offered, consisted of L1 and L2 GSL signers, L1 signer group including deaf, hard of hearing or hearing signers that acquired GSL from their immediate family environment from early childhood, and L2 signer group including deaf, hard of hearing or hearing signers that acquired GSL via educational procedures (Costello et al., 2006). L1 signers where not further defined as deaf or codas.

Due to General Data Protection Regulation (GDPR) issues and research ethics guidelines and regulations, responding to the questionnaire was anonymous, while participants personal information was restricted to a minimal set of metadata concerning demographic information on gender, age group, education level and GSL manner of acquisition (L1 vs L2).

Within a three-week period, the questionnaire was distributed among members of the GSL Community including deaf clubs and educational institutions. 91 distinct IP addresses were identified as having visited the questionnaire. However, only 32 participants completed the questionnaire, of which 17 identified themselves as L1 signers and 15 as L2 signers. Thus, only the responses of those 32 participants were considered in the analysis of the results. All participants were adults between 18 and 61 years of age.

3.1 Overview of the Results from the Pilot On-Line Survey

Participants were asked to rate the performance of each avatar in each signing occurrence in a 5-scale rating (Bad / Rather Bad / Average / Good / Very Good). The relative frequency distributions on this 5-scale rating for parts A and B are illustrated in the bubble charts of Figure **2** and Figure **3**, respectively, where the size of each bubble denotes the percentage of responses for a specific rating. In Part A, about 76% of the participants considered PAULA's performance "good" or "very good", while only 6% assign a low rate of "bad" or "rather bad". The relative frequencies for FRANÇOISE were 41% for good/very good ratings and 20% for bad/very bad ratings (Figure **2**).
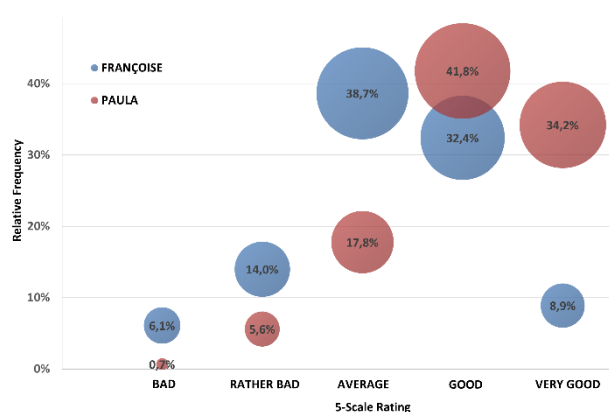
Figure 2: Relative frequency distribution of 5-scale rating for all signing occurrences of Part A.

The frequency distributions of Part B were similar as can be seen in Figure **3**. Obviously, the pairs of bubbles (for FRANÇOISE and PAULA at each rate) are either very close or significantly overlapped. For instance, about 36% and 39% of the rates were at the level "good" for PAULA and FRANÇOISE, respectively. Moreover, the relative frequencies at "rather bad" and "average" are also comparable. This similarity would become apparent if one drew lines that connect the centers of the bubbles for each avatar.

Considering the data from both parts, the descriptive analysis of results shows an overall preference for PAULA avatar performance. However, our goal is to investigate the preferences that the two sub-groups (L1 and L2) expressed towards the two avatars. Even though the collected metadata were based on participants' statements (e.g., they identified themselves as L1 or L2 signers), we strongly believe that nobody would benefit from misleading us given that the evaluation's scope is to strengthen the constructive cooperation between researchers and the deaf community.

In this sense, we considered the L1 and L2 participants two independent groups. It is worth mentioning that we target signers who favor this cooperation such as the volunteers who participated. To this end, we consider the participants a sample of the targeted population. However, we are aware of the random sampling process, and we plan to adopt it in the next evaluation phase when many more participants will be involved.
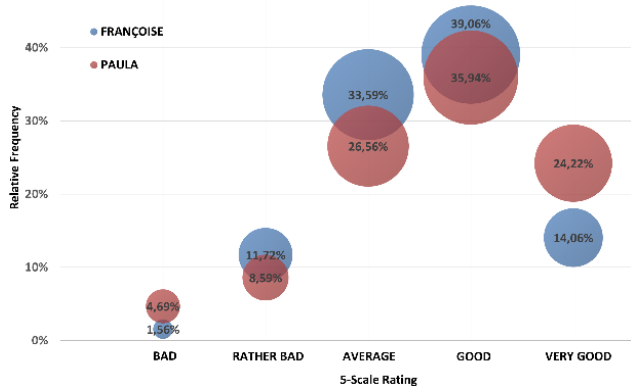

Figure 3: Relative frequency distribution of 5-scale rating for all signing occurrences of Part B.

3.1.1 Preferences Investigation of L1 and L2 Signers

In the light of the above, we hypothesized that the two subgroups expressed the same preferences towards the two avatars (NULL hypothesis). To explore this hypothesis, we conducted Mann Whitney U Tests[2] to test if there is a statistically significant difference in the rating of an avatar between the two groups. We applied the analysis on both PAULA and FRANÇOISE for Part A and Part B.

Figure 4: Part A: Distribution of relative frequencies of the 5-scale rating in Part A for both avatars in the two sub-groups (L1 and L2 signers).

The distribution of relative frequencies on the 5-scale rating for both avatars in the context of parts A and B for each group are illustrated in Figure **4** and Figure **5** respectively. Based on these results, one could observe (by comparing the first two columns of Figure **4**) that the results

for PAULA in Part A are very similar for both groups of signers. Although it seems that there are differences in the other cases i.e., rates of L1 and L2 groups for FRANÇOISE in Part A (3rd and 4th columns of Figure **4**), and Part B (3rd and 4th columns of Figure **5**), and for PAULA (first two columns of Figure **5**), the statistically significant ones, will be concluded by inferential statistics.

In Part A, the comparison of the two signer sub-groups (L1 vs L2) for both avatars resulted in the following:

PAULA: the resulted p-value was $0.17 > 0.05$ (the selected significance level), hence the NULL hypothesis cannot be rejected which can be interpreted that both sub-groups rate PAULA's response similarly.

FRANÇOISE: the resulted p-value was $0.0004 < 0.05$ and thus we can accept the alternative hypothesis and state that there is a statistically significant difference between the rates provided by L1 and L2 signers. Given that the median values of rates of each sub-group are equal to "AVERAGE" (i.e., percentages for "BAD", "RATHER BAD" and "AVERAGE" sum up to more than 50% in both groups of green shades), we cannot decide which sub-group provides higher rates to FRANÇOISE. However, by observing the modes of each sub-group (i.e., 43,34% of L1 and 37,89% of L2 rated this avatar of "AVERAGE" and "GOOD" performance respectively) we could say that L2 signers graded FRANÇOISE higher than what L1 signers did.
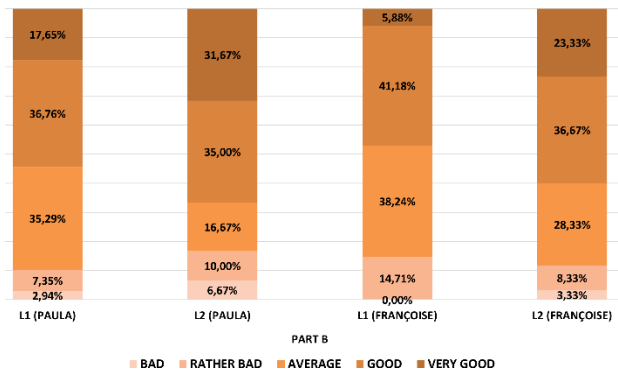

Figure 5: Distribution of relative frequencies of the 5-scale rating in Part B for both avatars in the two sub-groups (L1 and L2 signers).

In Part B, the comparison of the two signer sub-groups in respect to their ratings of the two avatars provides the following results:

PAULA: the resulted p-value was $0.087 > 0.05$, hence the NULL hypothesis cannot be rejected which can be interpreted that both sub-groups rate PAULA's response similarly. It is worth mentioning that this conclusion was also conducted for this avatar in Part A.

FRANÇOISE: the resulted p-value was $0.022 < 0.05$ and thus we can accept the alternative hypothesis and state that there is a statistically significant difference between the rates provided by L1 and L2 signers. In Part B (see last two columns of Figure **5**), the median values of rates of L1 and L2 signers are "AVERAGE" and "GOOD" respectively (i.e., percentages for "BAD", "RATHER BAD" and

---

[2] The Mann Whitney U Test is the alternative of the independent t-test. It is a non-parametric test proper for statistical analysis when the data are ordinal and there is no assumption of the distribution of the population and the two groups have unequal sizes.

"AVERAGE" sum up to more than 50% in L1 group, while "GOOD" is required to be included in L2 group), and thus we could say that again the L2 signers sub-group graded FRANÇOISE slightly higher than what L1 signers did.

3.1.2 Interpretation of Results with Respect to End-User Preferences

Regarding the first part (Part A) of the survey and the head-to-head presentation of the two avatars, for which participants were asked to choose the avatar that had a signing performance closer to the performance of a human, results showed that PAULA was the avatar of preference, as shown from the ratings as well as from the responses count from the Head-to-Head comparison; out of the total 608 signing occurrences (19 stimuli of images and videos multiplied by 32 participants), Paula was chosen in 428 of them.

The statistical analysis showed that the most frequent response for the totality of the signing occurrences for PAULA is "Good" and for FRANÇOISE is "Average" (Figure **2**). This finding is consistent with the obtained results from the head-to-head avatar comparison.

Even though a larger amount of data is necessary to safely draw conclusions, the here attempted interpretation of the results simply highlights the general tendency which favors PAULA's signing over FRANÇOISE's one.

In the second part of the questionnaire (Part B), each avatar was individually rated for its signing performance with respect to a compilation of signing occurrences consisting of isolated lemmas and phrases. The overall inspection of the collected data for Part B attests that both avatars performed equally well. An investigation of their performance with respect to the four movement parameters that were evaluated (hand movement, body movement, head and eye movement, mouth movement) led to the following findings: PAULA received higher rankings for hand movement and eyes movement, while FRANÇOISE was preferred over PAULA for her mouth movement. Both avatars were equally evaluated with respect to their body movement. These are important findings that need to be investigated in more signing occurrences, within context as well as in isolated instantiations.

With respect to the preferences comparison of two sub-groups among the GSL signers we comment on the following: In both parts of the questionnaire the two sub-groups expressed the same preferences regarding PAULA. However, the difference between them regarding FRANÇOISE's rating is a finding worth further investigating. Further interpretation of this finding given the collected data yields two additional insights; a good avatar performance is rated similarly by both groups of signers, L1 and L2. However, an average signing performance gives room for varying ratings among signers. In this case L1 signers are shown to be more consistent in their ratings than the group of L2 signers who participated in the survey. To be able to interpret these results it is important to redress this issue in the follow-up surveys.

For this pilot implementation of the on-line survey, the number of participants was sufficient to perform an initial descriptive analysis. However, to further investigate the participants' choices and their respective ratings with respect to different variables (i.e., gender, age, SL manner of acquisition (L1 vs L2), educational status etc), we need to extend our survey aiming at a broader randomly selected pool of participants.

4. Discussion

The reported findings from the pilot on-line survey on avatar performance evaluation provided significant feedback not only with respect to the targeted aspects of avatar performance, but also regarding methodological issues such as the outreach of the survey so that statistical analysis of results is better supported, various distribution issues among participant groups, the size and structure of the survey content and the phrasing of the requested tasks. This feedback is exploited in the user evaluation surveys the design of which is reported here. These will constantly address different SLs in the framework of our strategy of ongoing signer consultation on avatar development as implemented within the EASIER project.

The pilot implementation of the on-line survey has demonstrated a successful user-centered design and incorporates accessibility features of the shell questionnaire environment which may effectively achieve to engage signers in the development of signing avatar technology.

Planned accommodation of content from four SLs (GSL, LSF, DGS and DSGS) will enable a wide application of the questionnaire in the next period, which will provide significant input from the part of users regarding how they perceive the parameters of naturalness and comprehensibility of the synthetic signing and will further guide development of the EASIER avatar.

5. Acknowledgements

6. Bibliographical References

Costello, B., Fernandez, F., & Landa, A. (2006). The non-(existent) native signer: Sign Language research in a small deaf population. In B. Costello, J. Fernández, & A. Landa (Eds.), *Theoretical issues in Sign Language Research (TISLR) 9 Conference*. Florianopolis, Brazil.

European Union of the Deaf. (2018,). Accessibility of information and communication. https://www.eud.eu/about-us/eud-position-paper/accessibility-information-andcommunication [Accessed October 26, 2018]

Jennings, V., Elliott, R., Kennaway, R. and Glauert, J. (2010). Requirements for a signing avatar. In the *Proceedings of the Workshop on Corpora & Sign Language Technologies (CSLT)*, Satellite workshop of the LREC 2010 Conference, Valetta, Malta, p. 33–136.

Kacorri, H., Huenerfauth, M., Ebling, S., Patel, K. & Willard, M. (2015). Demographic and Experiential Factors Influencing Acceptance of Sign Language Animation by Deaf Users. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '15)*. Association for Computing Machinery, New York, NY, USA, 147-154.

https://doi.org/10.1145/2700648.2809860

Kipp, K., Nguyen, Q., Heloir, A. & Matthes, S., (2011). Assessing the deaf user perspective on sign language avatars. In The *Proceedings of the 13th International ACM SIGACCESS Conference on Computers and accessibility (ASSETS '11)*. Association for Computing Machinery, New York, NY, USA, p. 107–114. DOI: https://doi.org/10.1145/2049536.2049557

McDonald, J., Wolfe, R., Moncrief, R. and Baowidan, S. (2016). A computational model of role shift to support the synthesis of signed language. In the *Proceedings of the 12th Theoretical Issues in Sign Language Research (TISLR)*, Melbourne, Australia, p. 4–7.

Wolfe, R., McDonald, J. and Schnepp, J. C. (2011). An Avatar to Depict Sign Language: Building from Reusable Hand Animation. In *Proceedings of the International workshop on Sign Language Translation & Avatar Technology Workshop (SLTAT'11)*.

Wolfe, R., McDonald, J., Efthimiou, E., Fontinea, E-S., Picron, F., Van Landuyt, D., Sioen, T., Braffort, A., Filhol, M., Ebling, S., Hanke, T. and Krausneker, V. (2021). The Myth of Signing Avatars (long paper). In the *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, Shterionov D., Editor, Association for Machine Translation in the Americas (2021). p: 33-42.

World Federation of the Deaf. (2018). WFD and WASLI statement of use of signing avatars. https://wf-deaf.org/news/resources/wfd-wasli-statement-use-signing-avatars/ [Accessed March 14, 2018].