Online Evaluation of Text-to-sign Translation by Deaf End Users: Some Methodological Recommendations

Floris Roelofsen f.roelofsen@uva.nl
Lyke Esselink l.d.esselink@uva.nl
Shani Mende-Gillings s.e.mendegillings@uva.nl
University of Amsterdam, the Netherlands

Maartje de Meulder maartje.demeulder@hu.nl
Nienke Sijm nienke.sijm@hu.nl
Utrecht University for Applied Sciences, the Netherlands

Anika Smeijers a.s.smeijers@amsterdamumc.nl
Amsterdam University Medical Centre, the Netherlands

Abstract

We present a number of methodological recommendations concerning the online evaluation of avatars for text-to-sign translation, focusing on the structure, format and length of the questionnaire, as well as methods for eliciting and faithfully transcribing responses.

1 Introduction

There is no generally accepted methodology for evaluating the comprehensibility of avatars for text-to-sign translation, let alone for doing so *online*. Evaluation procedures designed in previous work generally involve on-site interaction between experimenters and participants (Gibet et al. 2011; Smith and Nolan 2016; Ebling and Glauert 2016; David and Bouillon 2018; Huenerfauth 2006; Kacorri et al. 2015, though see Quandt et al. 2021 and Schnepp et al. 2011 for exceptions). The COVID-19 pandemic has made it necessary to turn to online procedures, which come with additional methodological challenges. On the bright side, such online procedures, if effective, may also have benefits in a post-COVID-19 world.

We report work in progress on the evaluation of a recently developed prototype system for translating sentences that frequently occur in a healthcare setting, particularly ones that are used in the diagnosis and treatment of COVID-19, from Dutch into Dutch Sign Language (NGT). The system itself is described in some detail in Roelofsen et al. (2021). Here, we share some of the lessons we have learned in designing a methodology for evaluating this system online. Some of these lessons specifically concern the online nature of the evaluation procedure, but others are more general and would apply to on-site evaluation as well.

In the process of designing our methodology, we held a feedback session with seven deaf researchers at various career stages, all users of NGT and familiar with (socio-)linguistic experimental methodologies, in which we discussed a preliminary setup of the evaluation procedure. After incorporating feedback from this session we carried out a pilot study with five participants (all consider NGT (one of) their mother tongue(s)). While the feedback session had already led to important improvements of the design, the pilot study brought out a number of

further methodological issues, serious enough to render the results essentially uninterpretable. To address these issues, we have further adapted the design of the evaluation procedure, which is described in more detail in Section 4. The adapted procedure is already in use. Although it is too early to present quantified results, it is clear that the methodological adjustments we made are effective, as the issues experienced in the pilot study are no longer present. By sharing the lessons we have learned from the feedback session and the pilot study, we hope that other researchers evaluating avatars for text-to-sign translation in the future, be it online or on-site, will be able to avoid making the same mistakes we did initially and arrive at a suitable evaluation procedure more directly.

The extended abstract is organised as follows: Section 2 outlines the goals of our evaluation procedure, Section 3 discusses the design of the questionnaire, Section 4 turns to issues concerning elicitation and transcription of participants' responses, and Section 5 concludes.

2 Evaluation goals

As mentioned above, the system we are evaluating translates sentences that frequently occur in a healthcare setting, especially in the diagnosis and treatment of COVID-19, from Dutch to NGT. For instance, a healthcare professional may enter the sentence 'Gebruikt u medicijnen?' ('Do you use any medications?') and the system will produce a translation in NGT. Some translations have been pre-recorded on video, others are displayed by means of an avatar, making use of the JASigning avatar software (Kennaway et al., 2007; Ebling and Glauert, 2016). We are mainly interested at this point in evaluating the comprehensibility of these avatar translations.

More specifically, our primary goal currently is to answer the following three questions:

1. **Individual sign recognition**: To what extent do deaf NGT users recognise the individual signs that the avatar translations consist of?
2. **Sentence comprehension**: To what extent do deaf NGT users understand the avatar translations as intended at sentence level?
3. **Clarity**: How clear are the avatar translations that the system produces?

Measuring individual sign recognition alongside sentence comprehension provides us with additional insights as to *why* a sentence is (mis)understood, and highlights specific areas for improvement. For example, some participants may recognise individual signs yet misidentify the meaning of a sentence (or vice versa).

A secondary goal (equally important in general, but less central in the present study) is to find out how members of the deaf community in the Netherlands view avatar technology for sign language translation, and the potential application of such technology in various domains (cf., David and Bouillon 2018; Bouillon et al. 2021; Quandt et al. 2021, among others).

3 Design of the questionnaire

We will comment on three design features of the questionnaire: its structure, format, and length.

Structure In evaluating the comprehensibility of avatar translations, it is crucial to have a standard of comparison. Suppose, for instance, that we find that users correctly recognise 75% of the individual signs that the avatar produces. This information in itself does not tell us much. Is this a positive result, or a negative one? We cannot tell as long as we do not have a baseline. This concern is particularly relevant here for two reasons. First, some of the translations involve medical terms (e.g., 'intravenous drip') which may not be familiar to all participants and therefore poorly recognised even if they are signed correctly by the avatar. Second, there is considerable regional and intergenerational variation in NGT, which means that certain signs may be familiar to NGT users from one region/generation, but not to users from another. To address

this issue, we compare the comprehensibility of avatar translations to that of video recordings of the same sentences signed by a deaf signer. The core of the questionnaire, then, consists of two parts: one that assesses the comprehensibility of avatar translations, and one that does the same for baseline videos signed by a deaf signer. The avatar part precedes the baseline part, to avoid a learning effect when assessing the avatar.

After these two core comprehension parts, the questionnaire inquires about the participant's general perception of avatar technology for sign language translation, and their views on the potential application of such technology in various domains.

In addition, the questionnaire includes an introductory part (informed consent, information about the structure of the questionnaire, and some questions about the language background of the participants) and a closing part which checks whether the questions in the questionnaire were clearly posed and could be responded to in a satisfactory way.

Format All questions and instructions in the questionnaire are presented both in NGT (by means of pre-recorded videos) and in written Dutch. The person giving instructions and asking questions in the videos is a deaf NGT user, distinct from the signer in the baseline video translations discussed above. Participants are given a choice as to whether they want to watch the questions/instructions in NGT, read the Dutch text, or both. Most participants preferred the videos, but some chose to read. Several participants explicitly commented that they appreciated having this choice. Some explicitly commented that they found it pleasant that the person in the video was a deaf signer. All participants reported that the questions and instructions were clear.

Length We aim to keep sessions under 45 minutes to avoid concentration difficulties. This seems to work well—participants appear to be focused all the way through. What we have learned, however, is that this means that the number of test sentences has to be kept quite low. Our initial plan was to present 24 avatar translations and 24 corresponding baseline videos, but this turned out not to be feasible at all. We now present 12 avatar translations and 12 baseline videos, and this generally fits the 45 minute window.

Another lesson we learned is that, in order to measure the extent to which the individual signs in a sentence are correctly recognised, the length of test sentences should be restricted to around 7 signs. It is well-known that most adults cannot store more than 7 items in their short-term memory (Miller, 1956). Indeed, when we presented longer sentences in our pilot study and asked participants to list the individual signs in these sentences, they had great trouble reproducing the right sequence even if they had fully understood the meaning of the sentence as a whole. Since our aim here is not to test participants' short term memory capacity but just comprehension, we have decided to keep all test sentences relatively short (4-7 signs). In the evaluation sessions we are currently running this appears to work well.

4 Eliciting and transcribing responses

For a proper evaluation procedure (ensuring that responses are correctly understood by all parties), the responses that participants provide in NGT have to be simultaneously interpreted into Dutch. This is not straightforward if, as in our case, the experimenters are not fluent signers: one is a new signer using NGT on a daily basis and the other has taken a number of NGT courses but does not use the language daily. The online setting makes this issue even more acute. We are addressing this issue as follows. During an evaluation session, the participant does not open the questionnaire on their own computer. Rather, one of the experimenters opens the questionnaire on their computer and shares their screen. An experienced sign language interpreter, with high awareness of regional and generational variation, is present as well. Before getting started, we make sure that both the questionnaire and the sign language interpreter are visible for the participant. Participants answer questions in NGT, i.e., they do not need to type anything themselves.

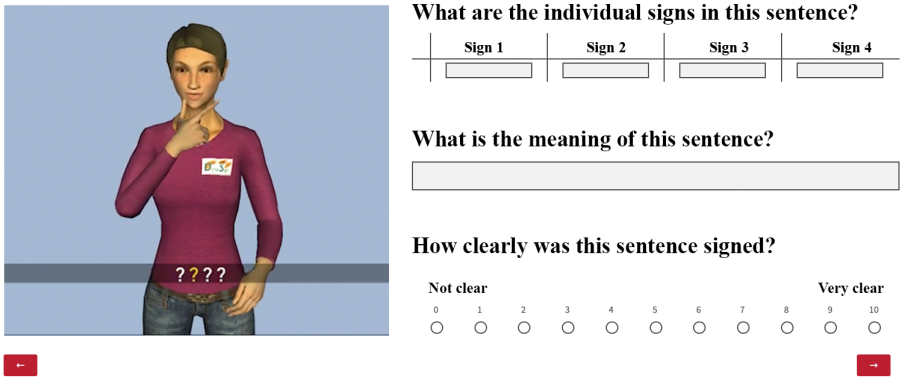

Figure 1: An example of an item in the questionnaire assessing comprehension of the avatar. For illustration, the questions are formulated in English here; in reality they are in Dutch and accompanied by instructions in NGT.

The interpreter interprets the answers into Dutch, one of the experimenters types the verbatim interpretation visible for the participants, so that they can check that their responses are properly interpreted. Typically, participants correct interpretations a few times each session and the transcript is then changed accordingly. In other cases, participants typically indicate explicitly that the interpretation is correct (usually with a confirming head nod after the transcription appears on the screen).

Finally, we turn to the issue of how to properly assess the extent to which participants recognise the *individual signs* in the avatar translations. This issue is more specific than the ones discussed above, but needs to be carefully addressed in any study that evaluates the comprehension of signing avatars. Indeed, the data obtained in our pilot study was uninterpretable mostly because we had not addressed this issue carefully enough.

In the pilot study, we gave participants instructions (both in NGT and in written Dutch) that they would be shown a video of an avatar signing a sentence and would then be asked three questions (i) What are the individual signs in the sentence? (ii) What is the meaning of the sentence as a whole? and (iii) How clearly was the sentence signed? Next, we showed participants a video, and then questions (i)-(iii), in Dutch. Responses to the first two questions (individual signs and sentence meaning) had to be entered in a textfield, while responses to the third question (clarity) had to be given on a scale from 0 to 10. The problem was that participants generally (with very few exceptions in fact) immediately started answering the second question. It was not sufficiently clear what was intended with the first question.

We took two measures to address this issue. First, rather than a single textfield for listing the individual signs in the sentence, we now present a separate textfield for each sign and label these textfields as 'Sign 1', 'Sign 2', etc (see Figure 1). Second, when giving instructions beforehand we now present two examples: one of an avatar translation with 'gloss subtitles', where the item in the gloss that corresponds to the current sign gets highlighted in yellow, and a second example of an avatar translation with question marks in the subtitles (see Figure 1). During the first sign the first question mark is highlighted, during the second sign the second question mark etc. Together with this second example video we also show the first two questions (concerning individual signs and sentence meaning, respectively), and exemplify what a possible response could look like. The question mark subtitles are also included in the actual test items. These two revisions of the design appear to achieve the intended effect: in the evaluation procedure we are currently running participants so far respond to all questions as intended.

5 Conclusion

In this extended abstract, we have shared a number of methodological lessons we have learned in designing and piloting an online procedure to evaluate the comprehensibility of an avatar for text-to-sign translation. We hope that the recommendations we have made concerning the structure, format, and length of the questionnaire and test items, as well as the elicitation and transcription of responses will be helpful for other researchers in designing their evaluation procedures. In the long run, we hope that they contribute to the development of more standardised methodologies and best practices for the evaluation of sign language technology.

Acknowledgments

We want to express our thanks to the deaf researchers that joined the feedback session for their insights. We gratefully acknowledge financial support from the Netherlands Organisation for Innovation in Healthcare (ZonMw, grant number 10430042010027) and the European Research Council (ERC, grant number 680220).

References

Bouillon, P., David, B., Strasly, I., and Spechbach, H. (2021). A speech translation system for medical dialogue in sign language—Questionnaire on user perspective of videos and the use of Avatar Technology. In *Proceedings of the 3rd Swiss Conference on Barrier-free Communication (BfC 2020)*, pages 46–54.

David, B. V. C. and Bouillon, P. (2018). Prototype of Automatic Translation to the Sign Language of French-speaking Belgium. Evaluation by the Deaf Community. *Modelling, Measurement and Control C*, 79(4):162–167.

Ebling, S. and Glauert, J. (2016). Building a Swiss German Sign Language avatar with JASigning and evaluating it among the Deaf community. *Universal Access in the Information Society*, 15(4):577–587.

Gibet, S., Courty, N., Duarte, K., and Naour, T. L. (2011). The SignCom system for data-driven animation of interactive virtual signers: Methodology and evaluation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(1):1–23.

Huenerfauth, M. (2006). *Generating American Sign Language classifier predicates for English-to-ASL machine translation*. PhD thesis, University of Pennsylvania.

Kacorri, H., Huenerfauth, M., Ebling, S., Patel, K., and 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*, pages 147–154.

Kennaway, R., Glauert, J., and Zwitserlood, I. (2007). Providing signed content on the internet by synthesized animation. *ACM Transactions on Computer-Human Interaction*, 14(3):1–29.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63:81–97.

Quandt, L. C., Willis, A., Schwenk, M., Weeks, K., and Ferster, R. (2021). Attitudes toward signing human avatars vary depending on hearing status, age of signed language exposure, and avatar type. Manuscript archived at PsyArXiv, June 25, doi:10.31234/osf.io/g2wuc.

Roelofsen, F., Esselink, L., Mende-Gillings, S., and Smeijers, A. (2021). Sign language translation in a healthcare setting. In *Translation and Interpreting Technology*.

Schnepp, J., Wolfe, R., Shiver, B., McDonald, J., and Toro, J. (2011). SignQUOTE: A remote testing facility for eliciting signed qualitative feedback. In *Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT)-2011*.

Smith, R. G. and Nolan, B. (2016). Emotional facial expressions in synthesised sign language avatars: a manual evaluation. *Universal Access in the Information Society*, 15(4):567–576.