Exploring automatic text-to-sign translation in a healthcare setting

Lyke Esselink[1,2,*], Floris Roelofsen[1], Jakub Dotlačil[3], Shani Mende-Gillings[1], Maartje de Meulder[4], Nienke Sijm[4], and Anika Smeijers[5]

[1]University of Amsterdam, [2]Radboud University Nijmegen, [3]Utrecht University

[4]University of Applied Sciences Utrecht, [5]Amsterdam University Medical Centre

[*]Corresponding author: l.d.esselink@uva.nl

April 21, 2022

Abstract

Communication between healthcare professionals and deaf patients has been particularly challenging during the COVID-19 pandemic. We have explored the possibility to automatically translate phrases that are frequently used in the diagnosis and treatment of hospital patients, in particular phrases related to COVID-19, from Dutch or English to Dutch Sign Language (NGT). The prototype system we developed displays translations either by means of pre-recorded videos featuring a deaf human signer (for a limited number of sentences) or by means of animations featuring a computer-generated signing avatar (for a larger, though still restricted number of sentences). We evaluated the comprehensibility of the signing avatar, as compared to the human signer. We found that, while individual signs are recognized correctly when signed by the avatar almost as frequently as when signed by a human, sentence comprehension rates and clarity scores for the avatar are substantially lower than for the human signer. We identify a number of concrete limitations of the JAsigning avatar engine that underlies our system. Namely, the engine currently does not offer sufficient control over mouth shapes, the relative speed and intensity of signs in a sentence (prosody), and transitions between signs. These limitations need to be overcome in future work for the engine to become usable in practice.

Keywords: access to healthcare information, sign language, avatar technology, user study

1 Introduction

Communication between healthcare professionals and deaf patients is challenging enough under normal circumstances (Fellinger et al., 2012), but has been especially difficult during the COVID-19 pandemic (McKee et al., 2020). Most healthcare professionals don't know the national sign language, COVID regulations often did not permit sign language interpreters to enter hospitals and clinics, interpreting via video relay is not always viable, and face masks conceal facial expressions and make lipreading impossible (Grote and Izagaren, 2020).

A survey among 179 deaf people in the Netherlands, carried out by one of the authors of the present article in January-March 2021, confirmed that the general inability of healthcare professionals to communicate in Dutch Sign Language (Nederlandse Gebarentaal, NGT) was perceived as

a very significant threat (Smeijers and Roelofsen, 2021). For instance, 88% of participants stated that they were worried about communication barriers should they need to be hospitalised with COVID-19, while, for comparison, only 33% stated that they were worried about the fact that friends and relatives would not be allowed to visit them in the hospital.

To address these concerns, we have explored the possibility to automatically translate phrases that are frequently used in the diagnosis and treatment of hospital patients, in particular phrases related to COVID-19, from Dutch or English to NGT. We developed a prototype system which displays translations either by means of pre-recorded videos featuring a deaf human signer (for a limited number of sentences) or by means of animations featuring a computer-generated signing avatar (for a larger, though still restricted number of sentences). We evaluated the comprehensibility of the signing avatar, as compared to the videos of a human signer.

We have concentrated on Dutch and English as the source languages for translation and NGT as the target sign language. The general problem we aim to address, however, is not specific to NGT but manifests itself for other sign languages as well. Therefore, we have aimed to design our prototype system in such a way that it could in principle be extended to include other source and target languages in a relatively straightforward way. In this respect, our system diverges from some existing text-to-sign translation systems, which are tailor-made for a specific target sign language and not easily portable to other languages (see Section 3.2 below). In particular, to our knowledge, none of the existing systems allows for translation from Dutch to NGT.

We should emphasise that a qualified human sign language interpreter is, whenever available, always to be preferred over a machine translation system, keeping in mind that even the use of sign language interpreters still has its own limitations (de Meulder and Haualand, 2021). We believe that it is worth investigating the extent to which a machine translation system can be of help in situations in which a human interpreter cannot be employed, including in certain medical settings where effective, instantaneous communication between healthcare professionals and patients can be of critical importance. But the aim of such technology should never be to replace human sign language interpreters across the board, and interpreting services of the highest possible quality remain critical in all domains.

The type of research reported here requires collaboration between researchers from several disciplines, bringing in different kinds of positionalities, knowledge and expertise. Before proceeding, we therefore include a brief note about the members of our research team and their respective contributions to the present project. De Meulder and Sijm are deaf; Esselink, Roelofsen and Smeijers are hearing new signers, with varying levels of proficiency in NGT; Dotlačil is hearing and not proficient in NGT. De Meulder is a scholar in Deaf Studies and applied linguistics, Sijm has a background in criminology and Deaf Studies. Esselink and Mende-Gillings have a background in Artificial Intelligence, Smeijers is a sign linguist and a medical doctor, Roelofsen has a background in linguistics and Artificial Intelligence, and Dotlačil contributed his expertise in statistical analysis. Esselink, Mende-Gillings, and Roelofsen designed and implemented the prototype system. Smeijers contributed her knowledge of the medical domain, and took main responsibility for the production of video translations used in the prototype system. The evaluation study was designed by Esselink, De Meulder, Roelofsen and Sijm, and was executed by Esselink and Roelofsen. The data from the study was analysed by Esselink, Roelofsen and Dotlačil.

The article is organised as follows. Section 2 provides relevant background information on sign languages and deaf communities, Section 3 discusses the prototype system we developed, Section 4 reports on the evaluation study, and Section 5 concludes.[1]

[1] A preliminary report on the prototype we developed was published as Roelofsen et al. (2021b).

2 Brief background on sign languages

Evidently, we cannot provide a comprehensive overview here of the linguistic properties of sign languages in general (see, e.g., Baker et al., 2016), nor of NGT in particular (see Klomp, 2021). We will, however, highlight some important features which any text-to-sign translation system needs to take into account.

First of all, sign languages have naturally evolved in deaf communities around the world (Kusters and Lucas, 2022). This means that, contrary to a rather common misconception, there is not a single, universal sign language used by all deaf people worldwide, but many different sign languages used on different scales by different deaf and hearing signers (Hou and de Vos, 2022).

Second, although sign languages exist in language ecologies in close contact with spoken languages, there is generally no direct correspondence between the sign language used in a given country and the spoken language used in that same country. For instance, while English is the mainstream spoken language both in the US and in the UK, American Sign Language (ASL) and British Sign Language (BSL) differ considerably from each other, as well as from spoken English. Such differences do not only pertain to the lexicon, but also to grammatical features such as word order. This means in particular that, to translate a sentence from English to ASL or BSL it does not suffice to translate every word in the sentence into the corresponding sign in ASL/BSL and then put these signs together in the same order as the words in the English sentence.

Third, for healthcare professionals to communicate exclusively through written text would not be satisfactory for most deaf patients. Deaf people have varying levels of access to auditory information, due to variance in hearing loss and differential access to education. Most deaf people, while they have developed skills in visual/tactile communication have no, reduced, or contextual sensory access to spoken languages. Contrary to popular belief, not all deaf people can lipread. In any case, lipreading is always mostly guesswork, where context is paramount. This can be a problem in medical settings with the use of medical jargon and words that are harder to anticipate. Moreover, health literacy has proven to be a barrier for many deaf patients (McKee et al., 2015; Napier and Kidd, 2013; Smeijers et al., 2011). In a medical setting it is critical to avoid miscommunication, to obtain reliable informed consent for interventions, and to foster an environment in which patients feel maximally safe. Relying exclusively on written text and lipreading will not achieve this.

Fourth, signs are generally not just articulated with the hands, but often also involve facial expressions and/or movements of the head, mouth, shoulders, or upper body. These are referred to as the *non-manual* components of a sign. A text-to-sign translation system has to take both manual and non-manual components of signs into account.

Fifth, related to the previous point, non-manual elements are not only part of the *lexical* make-up of many signs, but are also often used to convey certain *grammatical* information (comparable to intonation in spoken languages). For instance, raised eyebrows may indicate that a given sentence is a question rather than a statement, and a head shake often expresses negation. Such non-manual grammatical markers are typically 'supra-segmental', meaning that they do not co-occur with a single lexical sign but rather span across a sequence of signs in a sentence. Sign language linguists use so-called *glosses* to represent sign language utterances. For instance, the gloss in (1) represents the NGT translation of the question *Have you already eaten?*.

(1) $\overline{\text{YOU EAT ALREADY}}^{\text{brow raise}}$

Lexical signs are written in small-caps. They always involve a manual component and often non-manual components as well. The upper tier shows non-manual grammatical markers, and the horizontal line indicates the duration of these non-manual markers. In this case, 'brow raise' is used to indicate that the utterance is a question. A text-to-sign translation system should thus be able to integrate non-manual elements that convey grammatical information with manual and non-manual elements that belong to the lexical specification of the signs in a given sentence (Wolfe et al., 2011). This means that a system which translates sentences word by word, even if it re-orders the corresponding signs in accordance with the word order rules of the target sign language, will not be fully satisfactory. More flexibility is needed: word by word translation can be a first step, but the corresponding signs as specified in the lexicon, must generally be adapted when forming part of a sentence to incorporate non-manual markers carrying grammatical information.

3 A modular text-to-sign translation system

We have developed a prototype system which displays sign language translations either by means of pre-recorded videos featuring a deaf human signer, or by means of animations featuring a signing avatar. While video translations are clearly expected to be of higher quality, a translation system solely based on video translations would not scale up. With signing avatars, it may in principle be possible to build a system with much more comprehensive coverage. But how should appropriate avatar-based translations be generated, given the specific domain requirements? And are such translations comprehensible for the target end users, i.e., a widely varied group of deaf people? These questions have been the focus of our investigation. In what follows, we will therefore not say much about the video-based component of the system but concentrate mainly on the avatar-based component.

3.1 Sign synthesis

A crucial prerequisite for scalable automated text-to-sign translation is sign *synthesis*: the ability to create animations featuring a signing avatar. Broadly speaking, there are three ways in which this can be achieved: animation based on motion capture, keyframe animation, and scripted animation.

Motion capture makes it possible to obtain a library of high-quality animations for lexical signs, but requires expensive equipment and typically involves a lot of manual post-processing of the original data. Another major challenge under this approach is to modify the animations for lexical signs so as to incorporate non-manual grammatical markers (Courty and Gibet, 2010). Although it would in principle be possible to layer manual lexical animations with separate facial animations, exploring this option was not feasible within the timeframe of the present project and must be left for future work.

Keyframe animation results in lower quality lexical sign animations than motion capture. It does not require expensive equipment, but involves a lot of manual labour. Like motion capture, the problem of how to incorporate grammatical non-manual markers also applies to libraries of lexical signs obtained by means of keyframe animation.

The third synthesis method, scripted animation, offers a promising strategy to overcome this problem. On this approach, rather than directly animating each lexical sign, animations of lexical signs are generated *procedurally* based on structured specifications of the phonetic properties of these signs (Elliott et al., 2004). As in the case of keyframe animation, this also results in lower quality animations than could be obtained with motion capture techniques. However, no expensive

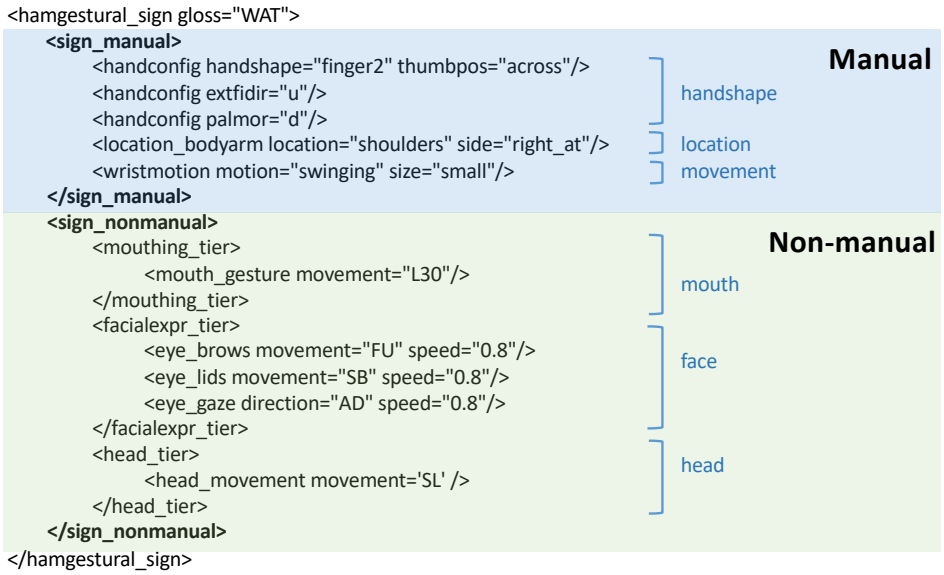


Figure 1: SiGML encoding of the NGT sign WAT ('what')

equipment is needed, and relatively little manual labour is required. The phonetic properties that make up the required specifications include (but are not limited to) the initial location, shape and orientation of the hands, possibly movements of the hands and other body parts, and facial expressions. Several formalisms have been developed to specify the phonetic properties of signs in a structured, computer-readable fashion (see Courty and Gibet 2010 for an overview). Arguably the most extensively developed and most widely used formalism is the Sign Gesture Markup Language (SiGML) (Elliott et al., 2004; Glauert and Elliott, 2011), which is based on the HamNoSys notation originally developed for the annotation of sign language corpora (Prillwitz et al., 1989; Hanke, 2004). For illustration, our SiGML encoding of the NGT sign WHAT is given in Figure 1. As can be seen in the figure, both manual features (handshape, location, movement) and non-manual features (mouth, face, head) are encoded.

SiGML specifications can be converted into animations by the JASigning avatar engine (Elliott et al., 2004; Kennaway et al., 2007; Jennings et al., 2010). This approach makes it possible, in principle, to integrate non-manual grammatical markers with the lexical signs that make up a sentence, although such functionality has not yet been thoroughly implemented in systems based on SiGML and JASigning to our knowledge.

Given these considerations, we opted to use SiGML and JASigning as a basis for sign language synthesis, and to implement a new functionality to automate the integration of non-manual grammatical markers with lexical signs. A basic library of SiGML specifications of around 2000 lexical signs in NGT was already compiled in the course of previous projects (Esselink 2020, see also Kennaway et al. 2007; Prins and Janssen 2014). While we have had to extend this library with healthcare-related as well as some general-purpose signs, the availability of an initial repertoire of signs encoded in SiGML was essential for a timely development of the system.

3.2 Translation

We now turn our attention from sign synthesis to the broader task of text-to-sign translation. Two approaches to this task can be distinguished, differing mainly in the type of intermediate representation that is employed in going from text to sign.

In the first approach, which we will refer to as the **gloss approach**, a given input sentence is transformed into a gloss of the corresponding sign language utterance. Next, based on this gloss representation, an avatar animation is generated.

(2) **Gloss approach**: text $\Longrightarrow$ gloss $\Longrightarrow$ animation

This approach is taken, for instance, by HandTalk, a Brazilian company that provides an automated text-to-sign translation service with Brazilian Portuguese and English as possible source languages, and American Sign Language (ASL) as well as Brazilian Sign Language (Libras) as possible target languages. HandTalk uses machine learning techniques to map input texts to the corresponding glosses, and a combination of keyframe animation and motion capture techniques to generate animations based on a given gloss.

In the second approach, which we refer to as the **phonetic approach**, the given input sentence is transformed into a sequence of phonetic representations of signs. Next, based on these phonetic representations, an animation is generated.

(3) **Phonetic approach**: text $\Longrightarrow$ phonetic representation $\Longrightarrow$ animation

This approach has been taken in work based on SiGML and JASigning (see, e.g., Zwitserlood, 2005; Kennaway et al., 2007; Prins and Janssen, 2014; Ebling and Glauert, 2016; David and Bouillon, 2018). Unlike in the gloss approach, applying machine learning techniques to carry out the first step, from text to phonetic representations, is not feasible because it would require the availability of large parallel corpora of texts and the corresponding phonetic sign representations, which do not exist and would be very costly to create. The process of manually generating phonetic representations requires expert knowledge of SiGML or a similar formalism. Rayner et al. (2016) have created a framework to ease this process, which is especially helpful if the sentences that need to be translated are all variations of a limited set of templates. For instance, the framework has been used to develop an application for translating railway announcements (David and Bouillon, 2018).

The gloss approach and the phonetic approach have complementary pros and cons. An advantage of the gloss approach is that it enables the use of machine learning technology to carry out the first part of the translation process. Disadvantages are that (i) the animation of each lexical sign involves substantial work, (ii) grammatical non-manual elements cannot be straightforwardly integrated with lexical signs, and (iii) all components of the system are tailor-made for a particular target sign language, i.e., no part of the system can be re-used when a new target language is considered. In particular, since no gloss-based system currently exists for NGT, this approach was not viable for our purposes.

Advantages of the phonetic approach are that (i) grammatical non-manual features can in principle be integrated with lexical signs (though this possibility remains largely unexplored) and (ii) part of the system, namely the software that generates avatar animations based on phonetic representations (i.e., JASigning or a similar avatar engine) is not language-specific and can in principle be re-used for any target sign language. The main disadvantage is that the initial step from text to phonetic representations involves a lot of manual work.

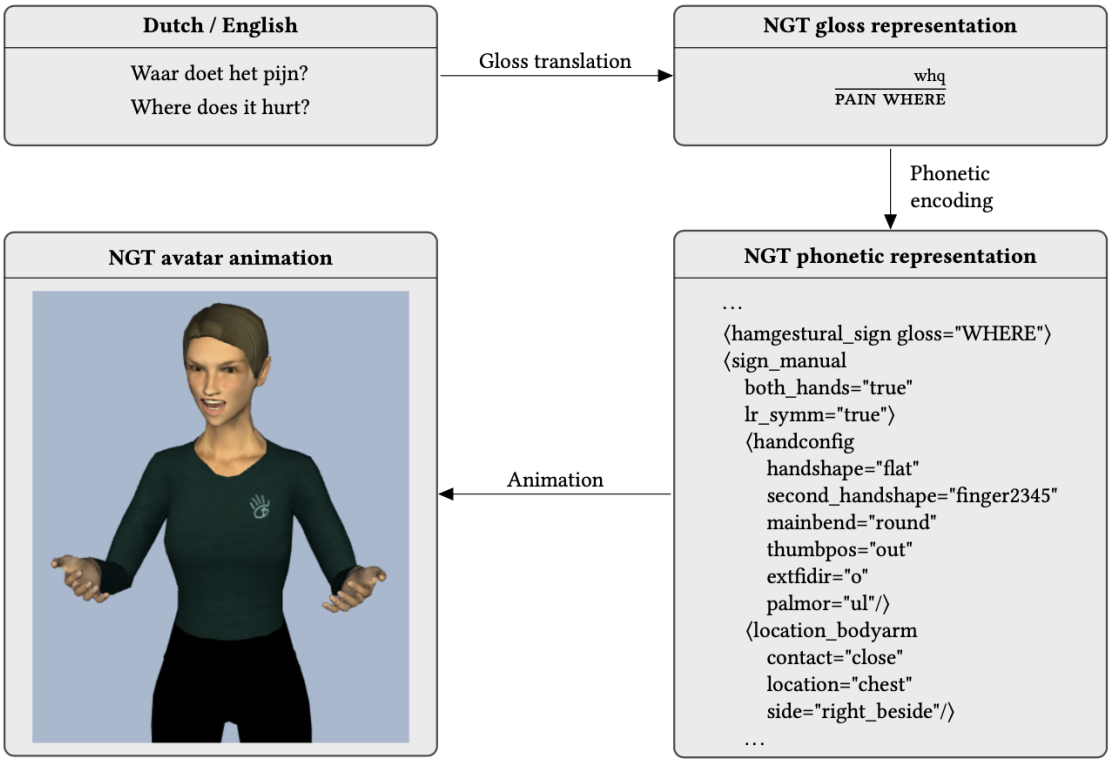


Figure 2: Overview of the modular translation pipeline

Given these considerations, we have taken a **modular approach**, which employs *both* a gloss representation *and* a phonetic representation in going from a given input text to an avatar animation of the corresponding sign language utterance. As depicted in Figure 2, our modular approach breaks the translation process up into three steps:

1. **Gloss translation**
   In this step, the given Dutch or English input sentence is mapped to a gloss representation of the corresponding NGT utterance. This can be done with a rule-based grammar or with machine learning, depending on use case requirements and availability of training data.

2. **Phonetic encoding**
   In this step, the NGT gloss is transformed into a computer-readable phonetic representation, in our case formulated in SiGML. This can be fully automated in a rule-based system, which can also integrate grammatical non-manuals.

3. **Animation**
   In this step, an avatar animation is generated based on the given phonetic representation. This procedure is not language specific, thus can be applied universally.

Consider, for instance, the Dutch/English input sentence in (4):

(4) Waar doet het pijn?
Where does it hurt?

The first step is to convert this sentence into the corresponding NGT gloss in (5), where 'whq' stands for the non-manual marking that is characteristic for constituent questions in NGT. While empirical studies have found quite some variation in the actual realisation of 'whq' in NGT (Coerts, 1992; de Vos et al., 2009), furrowed eyebrows are seen as the most canonical realisation (Klomp, 2021).

(5) $\overset{\text{whq}}{\overline{\text{PAIN WHERE}}}$

The second step is to map this gloss representation to a phonetic representation in SiGML, a fragment of which is displayed in Figure 2. Finally, this SiGML representation is fed into the JASigning avatar engine, which generates an animation (see Figure 3 on page 10 for a shapshot of the user interface of the system).

3.3 Implementation

Implementation choices depend on the specific use case requirements. Is it more important to achieve high precision, which a rule-based system allows, or to achieve broad coverage, which would favor an implementation involving machine learning? If the goal is to have optimal quality of lexical sign animations, one may opt to use motion capture, while scripted animation can be used in a scenario where scalability is of higher importance. The type and amount of resources available inevitably constrain one's choices as well. Is there enough data for machine learning? Is a rule-based grammar available for the given domain? Is motion-capture equipment available? What is the timeframe for development?

3.3.1 Use case requirements and implementation choices

Our main objective was to address the urgent concerns of deaf people in the Netherlands, ensuing from the COVID-19 pandemic, about the general inability of healthcare professionals to communicate in NGT (Smeijers and Roelofsen, 2021). Two specific requirements followed from this objective: (i) the system had to be developed within a short time-frame, and (ii) high accuracy of the delivered translations was more important than broad approximate coverage. In addition to these requirements, our implementation choices were also affected by the fact that resources were limited.[2]

Our aim has therefore *not* been to automate the entire translation process. In particular, automating the process of mapping input sentences to the corresponding NGT glosses using machine learning techniques would not have been feasible within a short time-frame, and would, even in the

[2]Since the timeframe and available resources for this research project were really quite different than for prototypical academic projects, we provide some details. Initial funding for the project was provided by an ad-hoc funding scheme set up by the Netherlands Organisation for Innovation in Healthcare (ZonMW) to address urgent COVID-related issues in the healthcare sector. The deadline for proposals in this funding scheme was two weeks after the call for proposals had been announced, funded projects had to start one month later, and had to be completed within six months, with a total budget of 25.000 euros. In this period, we designed and implemented the prototype system. Separate funding was used for the evaluation study.

somewhat longer term, most likely result in an unacceptably low accuracy rate for use in a healthcare setting. We therefore mainly focused on automating the phonetic encoding step, something that significantly reduces the manual labor needed in the overall translation pipeline. Automating the mapping from glosses to phonetic representations has not been done in previous work on NGT (Prins and Janssen, 2014) and, to the best of our knowledge, not in work on other sign languages either.

3.3.2 Selecting phrases for translation

We selected a set of phrases that are commonly used during the diagnosis and treatment of COVID-19, based on consultation with healthcare professionals at the Amsterdam University Medical Centre (AUMC) as well as direct experience (one of the authors is a medical doctor). We also consulted a list of phrases that was used in the SignTranslate system in the UK (Middleton et al., 2010).[3]

The resulting corpus was then divided into three categories: video-only, avatar-only, and hybrid. The first category, video-only, consisted mainly of sentences that could be divided into three further categories: emotional, complex, and informed consent. Sentences concerning the patient's emotional well-being require a high level of empathy to be conveyed, which is difficult to achieve in a satisfactory way with an avatar given the current state of the art. We therefore deemed that video translations were necessary for these sentences. Sentences were classified as complex when they involved a combination of several statements and/or questions, or required a demonstration of pictures or diagrams along with an explanation (see Appendix B, Figure 9 for an example). Finally, in the case of questions and statements concerning informed consent it is especially important to leave no room for potential misunderstandings. To ensure this, we chose to always offer video translations of these sentences.

The second category, avatar-only, consisted of sentences with many variations differing by only one word or phrase, indicating for instance the time of day or a number of weeks. It would not have been feasible to record a video translation for each version of these sentences.

The third category, hybrid, consisted of sentences that did not fall into one of the other two categories. For these, the system offers both a video translation and an avatar translation. In some cases, the avatar translation is slightly simplified compared to the video translation (e.g., some long sentences were broken up into several smaller ones).

After categorising all of the sentences, those from the first and third category were translated into NGT and recorded by a team consisting of a sign language interpreter and a deaf signer. Translations were checked by one of the authors (Smeijers), who is a sign linguist and a medical doctor. This resulted in a collection of 139 video translations. The sentences from the second and third category (including all variations) together comprised 7720 sentences for avatar translation.

3.3.3 Constructing phonetic representations

In order for the system to operate fast at run-time, we pre-processed all sentences and stored phonetic SiGML representations of their translations in a database. At run-time, the system only queries this database and does not compute any translations on the fly.

[3]The SignTranslate system was developed in the UK around 2010 to translate phrases common in a healthcare setting from English to British Sign Language. Translations were displayed by means of videos, not by avatar animations. Evidently, the system was not specifically targeted at COVID-19 healthcare. However, many general-purpose phrases are also relevant in the diagnosis and treatment of COVID-19.

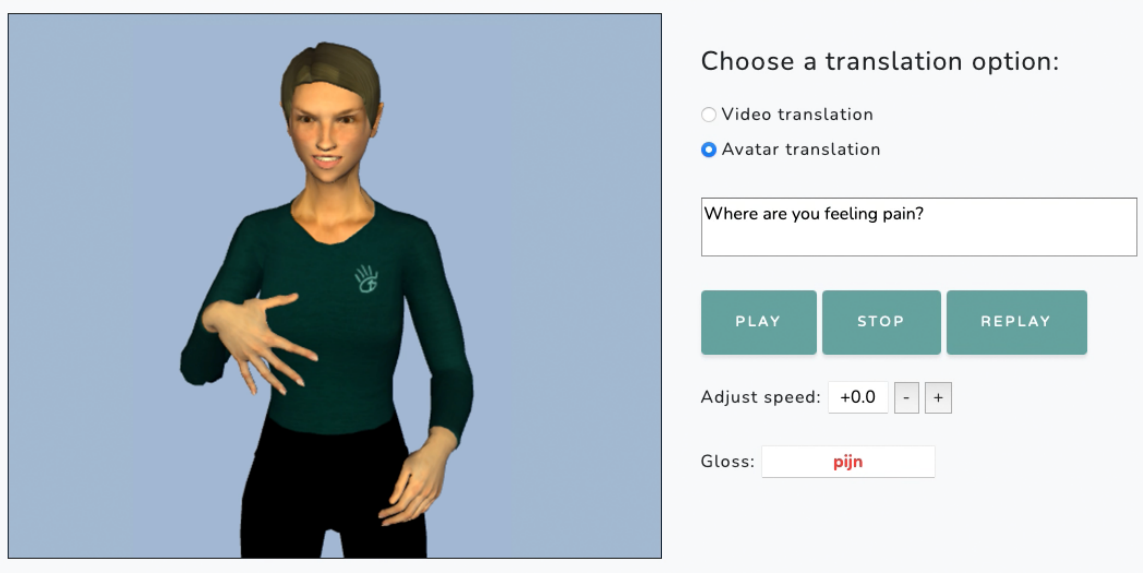


Figure 3: User interface of the system

To construct the SiGML representations of full sentences, we developed a system which, when given the gloss representation of a sentence in NGT, creates the SiGML code for that sentence. It first retrieves the SiGML code for all lexical signs in the given gloss from a lexical database, and then adapts this code to add non-manual grammatical elements. For instance, in the case of questions, the program makes sure that the sentence ends with PALMS-UP, a sign that can be used in NGT to mark questions, and adds raised eyebrows, both to PALMS-UP and to the sign preceding it.[4]

3.4 User Interface

We developed an online user interface (see Figure 3). Healthcare professionals can choose a translation format (video or avatar) and enter a sequence of search terms. Based on their input they are presented with a list of available sentences from the database. These sentences may differ depending on the translation format chosen (video/avatar). After selecting a sentence, the translation is offered in the chosen format.

As mentioned earlier, some of the possible input sentences differ only in one word or phrase. These sentences can be thought of as involving a general template with a variable that can take several values, such as a day of the week, a time of day, or a number of times / minutes / hours / days / weeks / months. When a user wants to translate such a sentence, they first select the template and then provide the intended value for the variable. For example, they may select the

[4] PALMS-UP and raised eyebrows are prototypical question markers in NGT, but questions can be marked in other ways as well (Coerts, 1992; de Vos et al., 2009). Furrowed eyebrows, for instance, are also sometimes used for this purpose. We always included PALMS-UP and raised eyebrows, but more research is needed to determine exactly under which conditions these question markers are used in NGT and under which conditions they are replaced by or combined with other markers.

template "I am going to explain more at *time*", and then select a particular time (as illustrated in Appendix B, Figure 10).

While JASigning in principle offers several avatars for sign language animation, there are differences in execution between these avatars. Our user interface therefore only makes use of one of them, Francoise (see Figure 3), and does not allow the user to choose different options. We intend to further optimise the visualisation of the avatar in future work.

4 Evaluation

To evaluate the implemented prototype system, we conducted an online survey among 22 deaf NGT users. 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 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.

4.1 Research questions

We focused on the following research questions:

RQ1. **Comprehensibility at the level of individual signs**
To what extent are individual signs understood as intended when performed by the avatar?

RQ2. **Comprehensibility at sentence level**
To what extent is the NGT translation of a given input sentence understood as intended when performed by the avatar?

RQ3. **Influence of interaction between participants and experimenters**
When evaluating the comprehensibility of a signing avatar in an online environment, how does interaction between participants and experimenters, or the lack thereof, affect the results?

The first two research questions concerned the comprehensibility of the implemented system. The third research question on the other hand concerns the methodology of evaluating signing avatars in an online setting. Of course, besides these three research questions there are other pertinent ones as well, concerning for instance the attitude of potential end-users towards signing avatars in general and towards our system in particular. We did include some questions in our survey to probe such attitudes, and will briefly report participants' responses to these questions below, but leave an in-depth investigation for a separate study.

4.2 Methodology

4.2.1 Participants

To recruit participants, we recorded a video in NGT which briefly explained our project and invited people to sign up as a participant of the evaluation study. This video was posted on various social

		Guided	Unguided	Total
Age group	18-30	0	6	6
	31-40	5	3	8
	41-50	2	2	4
	51+	2	2	4
Gender	Female	8	10	18
	Male	1	3	4
Region[1]	Central	2	4	6
	Northern	3	5	8
	Eastern	0	0	0
	Southern	0	1	1
	Western	4	3	7
Mother tongue	NGT	7	12	19
	Dutch	0	0	0
	NGT and Dutch	2	1	3
Frequency of NGT use[2]	Daily	9	11	20
	Regularly	0	2	2
	Occasionally	0	0	0
	Rarely	0	0	0
Frequency of use of interpreter	Daily	0	0	0
	Regularly	7	9	16
	Occasionally	1	4	5
	Rarely	1	0	1
Communication methods with non-signers in the absence of an interpreter	Lipreading	7	13	20
	Speech recognition	3	2	5
	Writing	8	7	15
	Signing	2	8	10
	Using voice	6	12	18
	Pointing	2	0	2
Frequency of use of other communication methods	Daily	3	4	7
	Regularly	4	6	10
	Occasionally	2	2	4
	Rarely	0	1	1

Table 1: Demographic information about the participants

[a]Regions are divided into the following provinces of the Netherlands: Central (Utrecht, Flevoland); Northern (Groningen, Friesland, Drenthe); Eastern (Gelderland, Overijssel); Southern (Brabant, Zeeland, Limburg); and Western (Noord-Holland, Zuid-Holland).

[b]Frequencies are defined as: daily (every day); regularly (a few days a week); occasionally (a few days a month); and rarely (a few days a year).

media platforms, including multiple Facebook groups, Instagram, and LinkedIn. We also asked personal contacts in the deaf community to distribute the video and ask their contacts to do the same. To sign up, interested people were asked to fill in a short form collecting their contact information and some demographic information. They were also asked whether they would prefer to participate in the online study with or without guidance of the experimenters. All these questions were asked both in written Dutch and in NGT through videos featuring a deaf signer.

We recruited 23 participants in total, but one of them did not complete the survey in the end, so the results of the survey are based on the responses of 22 participants. Table 1 provides an overview of their demographic information (the table contains aggregate data as well as data for both experimental groups, GUIDED and UNGUIDED, see Section 4.2.2 right below on the experimental design). Participants were spread across age groups relatively evenly. There were many more female (18) than male (4) participants. Most participants were from Central provinces (Utrecht, Flevoland), Northern provinces (Groningen, Friesland, Drenthe), or Western provinces (Noord-Holland, Zuid-Holland). Only one participant was from the Southern provinces (Brabant, Zeeland, Limburg), and no participants were from the Eastern provinces (Gelderland, Overijssel).

As for language background, 19 participants identified only NGT as their mother tongue, and 3 participants identified both Dutch and NGT as their mother tongue. No participants identified Dutch as their sole mother tongue. 20 participants indicated that they use NGT *daily*, while 2 participants used NGT *regularly* (a few days a week). No participants used NGT *occasionally* (a few days a month) or *rarely* (a few days a year). 16 participants indicated that they *regularly* make use of an interpreter, 5 indicated that they *occasionally* do, and 1 indicated that they *rarely* do.

When communicating with non-signers in the absence of an interpreter, participants indicate that they use a combination of various communication methods. The most frequent methods are lipreading (20 participants), speaking (18 participants), and writing (15 participants). 10 participants indicated that they continue to communicate through sign language / gestures in such situations. Less frequent methods include the use of speech-to-text recognition software (5 participants), and pointing at people/objects (2 participants).

Methods to communicate with non-signers in the absence of an interpreter are used *daily* by 7 participants, *regularly* by 10 participants, *occasionally* by 4 participants, and *rarely* by 1 participant.

We collected this rather detailed demographic information at the recruitment stage with the intention to create groups of participants that were optimally counter-balanced in terms of age group, region, and language background. However, because the total number of recruited participants was relatively low, we decided to include all of them in the survey.

4.2.2 Experimental Design

Survey We designed an online survey consisting of four parts.[5] Instructions and questions in the survey were all posed both in Dutch (text) and in NGT (videos) in order to be optimally accessible. At the beginning of the survey a general outline was provided, and participants were asked to give informed consent. All four parts of the survey started with an explanation of the task that had to be performed in that part of the survey. All sessions took place in June-July 2021.

Part 1 contained demographic questions about the participant (e.g. age group, gender, region) and their use of NGT (e.g. do they consider Dutch and/or NGT their mother tongue, how often do they use NGT, how often do they use an interpreter). Most of these questions were already asked

[5]Certain elements of the set-up and process were informed by a pilot study and feedback session with seven deaf researchers in linguistics and related fields (for more details see Roelofsen et al. 2021a).

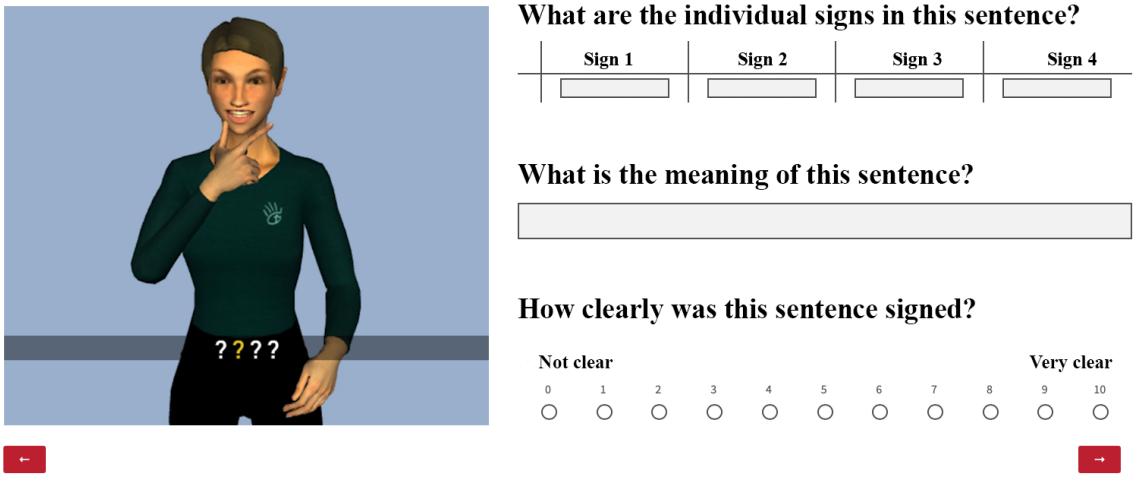


Figure 4: Example question in Part 2 of the survey

in the recruitment form, but were repeated here to ensure pseudonymization of the responses and informed consent from the participants for use of the data.

Part 2 assessed the comprehensibility of the avatar. It comprised 12 recordings of the avatar signing a sentence. For each sentence, the participant was asked to determine the meaning of each individual sign, the meaning of the sentence as a whole, and to rate how clearly the sentence was signed (see Figure 4). Participants were explicitly instructed that they were allowed to replay the recording as often as they wanted to, and that there was no time limit to perform the task. They were also shown an example item, to familiarize themselves with the appearance of the avatar and the format of the questions that they were supposed to answer.

Part 3 was similar to Part 2, only with video recordings of a human signer rather than an avatar. The sentences were the same as in Part 2, and participants were again asked to determine the meaning of each individual sign, as well as the meaning of the sentence as a whole, and to rate how clearly the sentence was signed, exactly as in Part 2.

Finally, *Part 4* consisted of general questions about the development of sign language avatars (e.g. did participants feel it was useful, in which scenarios could this technology be applied) and on the evaluation methodology (e.g. were the questions clear, was it easy to provide answers to the questions). For a complete overview of all the questions in the survey, with hyperlinks to the animations and videos that were used, see Appendix A.

Groups: guided vs unguided We divided participants into two groups, GUIDED ($n = 9$) and UNGUIDED ($n = 13$), based on the preference they had indicated in the recruitment form. Participants in the GUIDED group took the survey while in a conference call with two experimenters and a sign language interpreter.[6] During the conference call, the experimenters displayed the survey on their computer and shared their screen with the participants. Participants provided their answers

[6]The experimenters, Esselink and Roelofsen, were both hearing, with varying levels of proficiency in NGT. The sign language interpreter was highly experienced, familiar with many regional/generational variants of NGT, and was extensively briefed before the experiment.

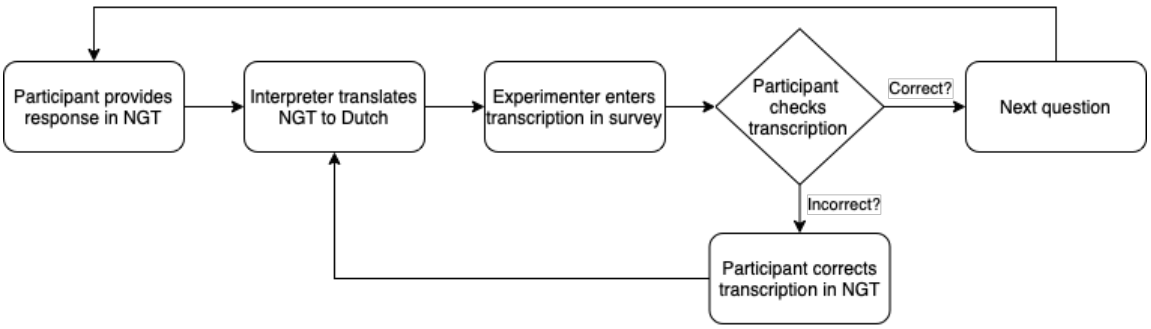


Figure 5: Feedback loop to ensure faithful textual transcription of responses provided in NGT

to the questions in the survey in NGT. To ensure a faithful textual transcription of these answers, the experimenters made use of a *feedback loop* which worked as follows (see Figure 5): first, the responses provided by the participants in NGT were translated to Dutch by the interpreter, and entered in the survey by one of the experimenters. Next, participants were given an opportunity to check and correct the textual transcription of their response before moving on to the next question. This feedback loop proved to be useful, as all participants corrected a transcription at least once.

Participants in the UNGUIDED group completed the survey without any guidance. They were sent the link to the survey via email, and were asked to complete it before a certain date. While these participants were able to view all instructions and questions in the survey in NGT through pre-recorded videos, they were unable to provide their responses in NGT. Instead, they entered Dutch text in the survey directly.

4.2.3 Coding comprehension data

For each sentence in Part 2 and 3, we encoded whether or not the participant correctly recognized each individual sign and interpreted the sentence as intended. In case the response was partly correct and partly incorrect we also encoded the type of error that was made. An overview of the different numeric labels we used to code the responses is provided in Table 2.

For individual signs, response types were coded as follows. Code 0 indicates that the participant

Code	Individual signs	Sentences
0	No response	No response
1	Wrong response provided	Wrong response provided
2	Manual component recognized correctly, but oral component not	Sentence radical recognized correctly, but sentence type not
3	Two possible interpretations provided, one of which correct	Two possible interpretations provided, one of which correct
4	Correct	Correct

Table 2: Encoding key

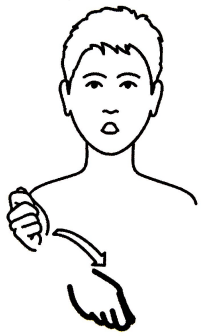


Figure 6: The manual component of the signs RESULT and PASSED in NGT. The two signs differ only in their oral component

did not provide a response. Code 1 indicates that an incorrect response was provided. Code 2 indicates that the participant correctly recognized the manual component of the sign, but did not correctly recognize the oral component (e.g. RESULT and PASSED share the same manual component, visualised in Figure 6, but have different oral components). Code 3 indicates that a participant provided two possible interpretations, one of which was correct. Finally, Code 4 indicates that the sign was correctly identified.

For entire sentences, response types were coded in a similar way. Again, Code 0 indicates that the participant did not provide a response at all. Code 1 indicates that the participant did not interpret the sentence correctly. Code 2 indicates that the participant correctly identified the meaning of the sentence radical, but confused the sentence type (e.g. the response was formulated as a question when the actual sentence was a statement). Code 3 indicates that the participant provided two possible interpretations, one of which was correct. Finally, Code 4 indicates that the sentence was correctly interpreted.

4.3 Results

4.3.1 Comprehension

Table 3a shows the proportion of each response type in percentages, for individual signs and sentences, split across the two signers (the avatar and the human signer). Note that 'partly correct' responses (coded as 2 or 3, depending on the type of error) occurred very infrequently. Therefore, in Table 3b, the results are presented in a more compact format, taking all incorrect responses (coded as 1, 2, or 3) to form a single category. Finally, in Table 3c, we go one step further and present the results in a binary format, distinguishing only between correct responses on the one hand and incorrect or missing responses on the other.

Comprehension rates for individual signs We zoom in now on Table 3b and first consider individual signs. Note that the comprehension rate of individual sings performed by the avatar (86.93%) is lower than the comprehension rate of individual signs performed by a human signer (92.28%), but both rates are quite high and the difference between them is rather small. Also note that, in the case of the human signer, almost all cases in which a sign was not correctly recognized

		Individual signs		*Sentences*	
Code	**Response type**	**Av**	**Hu**	**Av**	**Hu**
0	No response	6.32	6.85	4.17	0.00
1	Wrong response	4.40	0.72	18.56	1.89
2	Partly correct: A	2.27	0.00	1.89	0.76
3	Partly correct: B	0.07	0.14	0.76	0.00
4	Correct response	86.93	92.28	74.62	97.35

(a) Fine-grained

		Individual signs		*Sentences*	
Code	**Response type**	**Av**	**Hu**	**Av**	**Hu**
0	No response	6.32	6.85	4.17	0.00
1-3	Incorrect response	6.75	0.87	21.21	2.65
4	Correct response	86.93	92.28	74.62	97.35

(b) All incorrect responses aggregated

		Individual signs		*Sentences*	
Code	**Response type**	**Av**	**Hu**	**Av**	**Hu**
0-3	No correct response	13.07	7.72	25.38	2.65
4	Correct response	86.93	92.28	74.62	97.35

(c) All incorrect and absent responses aggregated

Table 3: Proportion of each response type in percentages, for individual signs and sentences, split across the two signers (**Av**atar vs **Hu**man)

are ones in which *no response* was given at all (6.85%). An incorrect response was only given 0.87% of the time. In the case of the avatar on the other hand, an incorrect response was given 6.75% of the time, while no response was given in 6.32% of the cases.

Comprehension rates for full sentences We now turn to the comprehension rates for full sentences. Here we see a bigger disparity between the avatar and the human signer. Participants were almost always able to correctly identify the meaning of a sentence when it was signed by a human (97.35%). This was not the case for sentences signed by the avatar: participants found it harder to correctly identify the meaning of these sentences, with a comprehension rate of 74.62%. They provided an incorrect response 21.21% of the time, and no response at all 4.17% of the time.

Most common mistakes We now take a closer look at which individual signs were not always recognized correctly. Table 4 lists all signs for which either no response was given more than twice or an incorrect response was given more than twice. For each of these signs, the Table indicates the number of responses of type 0 (no response at all), 1 (wrong response), and 2 (manual component recognized correctly, but oral component not) as a percentage of the total number of responses that were elicited for that sign performed by the given signer (avatar or human). The signs are divided into four categories: A (grammatical markers), B (pronouns), C (signs whose interpretation crucially relies not just on the manual component of the sign but also on mouthing), and D (miscellaneous).

		code 0		*code 1*		*code 2*	
Category	**Sign**	**Av**	**Hu**	**Av**	**Hu**	**Av**	**Hu**
A	INDEX	27.3	41.8	4.6	1.8	–	–
grammatical	PALMS-UP	8.2	16.7	9.1	1.5	–	–
B	YOU	–	11.9	–	–	–	–
pronouns	I	–	18.2	–	3.0	–	–
C	SOMETIMES	4.6	–	–	–	54.5	–
mouthing	FOR	22.7	–	–	–	36.4	–
	RESULT	–	–	–	–	22.7	–
	BACK	4.6	–	4.6	–	13.6	–
D	PAST	40.9	–	27.3	–	–	–
miscellaneous	SEVEN	50.0	–	13.6	–	–	–
	DAY	27.3	–	40.9	–	–	–
	INTRAVENOUS DRIP	31.8	–	9.1	–	–	–
	FAIL	9.1	–	27.3	–	–	–
	SORRY	4.6	–	13.6	–	–	–
	MORE	4.6	–	13.6	–	–	–

Table 4: Individual signs which were either not explicitly recognized more than twice or incorrectly recognized more than twice. In each case, we express the number of incorrect/missing responses as a percentage of the total number of responses elicited for the given sign performed by the given signer (**Av**atar vs **Hu**man)

The signs in category A are ones glossed as INDEX and ones glossed as PALMS-UP. INDEX signs are ones that refer back to something that has been introduced earlier in the same sentence. This 'doubling' mechanism is common in NGT but generally seems optional: leaving such signs out usually does not change the meaning of the sentence. In this sense, such signs are purely 'grammatical', they do not contribute any content. The PALMS-UP sign is used for various purposes in NGT. In the sentences under consideration, it was always used to mark a sentence as a question. When used for this purpose, PALMS-UP is generally optional as well. If it is left out, the signer's facial expression is generally sufficient to convey that she is asking a question rather than making a statement.

We see that these two signs were often not recognized explicitly (INDEX 27.3% of the time when signed by the avatar, and 41.8% of the time when signed by a human; PALMS-UP 8.2% of the time when signed by the avatar, and 16.7% of the time when signed by a human). This may well be connected to the fact that these signs are optional, and do not contribute any content to the meaning of the sentence as a whole that is not already conveyed by other elements as well. This may make them less salient, resulting in participants 'skipping over' them. They were also recognized incorrectly in some cases, especially when signed by the avatar (INDEX 4.6% of the time; PALMS-UP 9.1% of the time). A possible explanation for this result is that both signs have several other possible interpretations/grammatical functions as well. Indeed we find some of these among the interpretations provided by our participants (e.g., INDEX was sometimes interpreted as YOU or FOR, and PALMS-UP was sometimes misinterpreted as WHERE).

The signs in category B are the pronouns YOU and I. We see that these signs were quite often not explicitly recognized by our participants when signed by the human signer, though this did not occur with the avatar. A possible explanation for this is that these signs are often made very fast by human signers, and they are often co-articulated with adjacent signs. The avatar on the other hand, produced these signs at a slower pace, and did not co-articulate them with adjacent signs. When looking more closely at the data, we find that there were three sentences in which our participants failed to explicitly recognize these signs. Glosses of these sentences are given below.

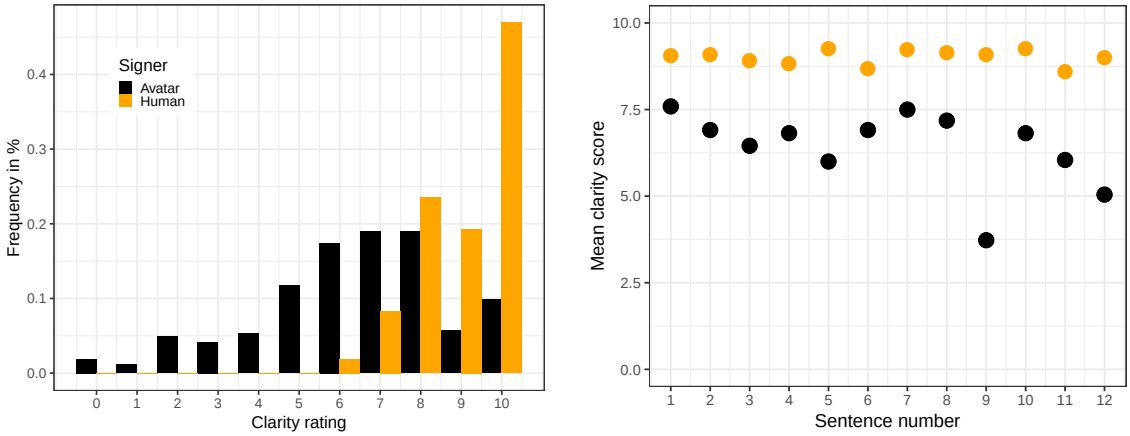
(6) SORRY I FAIL YOU INTRAVENOUS DRIP I COLLEAGUE CALL
not explicitly recognized: YOU (21 times), I (12 times)

(7) HEARING-AID OR COCHLEAR-IMPLANT HAVE YOU
not explicitly recognized: YOU (3 times)

(8) PAST SEVEN DAY YOU ALREADY CORONA TEST
not explicitly recognized: YOU (3 times)

The signs in category C are SOMETIMES, FOR, RESULT, and BACK. To recognize these signs correctly it is crucial to recognize not only their manual component but also the accompanying movement of the mouth. This is because for each of these signs there is at least one other sign with the same manual component but different mouthing.

We see that this is particularly problematic for the avatar. For instance, SOMETIMES was misinterpreted 54.5% of the time as MAYBE, which has the same manual component, and similarly, RESULT was misinterpreted 22.7% of the time as PASSED. Note that such misinterpretations did not arise when the signs were performed by a human signer. For the sign FOR, we see that, again only when performed by the avatar, it was misinterpreted 36.4% of the time and not explicitly recognized 22.7% of the time. A possible explanation for the fact that it was so often not explicitly recognized is that, when the mouthing is not correctly perceived, the sign can easily be mistaken



(a) Relative frequency of clarity scores

(b) Mean clarity scores per sentence

Figure 7: Clarity scores

for an INDEX sign, and we already saw that INDEX signs are often not explicitly recognized (possibly because they are optional and therefore perhaps less salient).

Category D contains seven miscellaneous signs, i.e., these are not grammatical markers, pointing signs, or signs whose recognition crucially relies on mouthing. Note that all signs in this category were correctly recognized when signed by a human signer, but not always when signed by the avatar. The first three signs, PAST SEVEN DAY were signed sequentially as one phrase (see example (8) above). When signed by the avatar, this sequence of signs was quite fast, with a high degree of co-articulation. This made it especially difficult to identify the sign SEVEN. An additional intricacy is that DAY is very similar to MONTH, differing only slightly in the location of the hand. This contributed to the frequent misinterpretation of the sign.

Turning now to the other three signs in category D, a possible explanation for the low comprehension rate of INTRAVENOUS DRIP when signed by the avatar is that it is, presumably, a rather infrequent sign. When performed by a human signer, however, it was always recognized correctly. The low comprehension rates of FAIL and MORE are possibly due to the fact that there are several other signs which are very similar and would have made sense in the given contexts (in particular, participants misinterpreted FAIL either as LOST or as WHERE, and MORE either as FIRST or as AGAIN). Finally, our participants were sometimes confused about the sign SORRY when performed by the avatar. The interpretations they provided in these cases were I, HEART, and CARD. Since the sign SORRY is actually quite different from I, HEART, and CARD, the only possible explanation we can offer for this confusion is that the avatar's rendering of the sign SORRY was not quite successful.

4.3.2 Clarity

Figure 7 provides an overview of the clarity scores participants gave to avatar animations generated by the system, compared to videos of a human signer. Recall that all participants saw avatar animations for 12 sentences, followed by videos of a human signer for the same 12 sentences. Clarity ratings were given on a scale from 0 to 10, where 0 was labeled as 'not clear' and 10 as 'very clear'.

Clarity scores given for avatar animations ranged between 0 and 10, with a mean of 6.4. Clarity

scores for videos ranged between 6 and 10, with a mean of 9.0. In Figure 7a we have plotted the relative frequency, in percentages, of each clarity score given to avatar animations and videos, respectively, summed across all participants. This figure shows that scores of 6, 7, and 8 were most frequently given to avatar animations, while scores of 8, 9, and 10 were most frequently given to videos of a human signer. Videos received a maximal score of 10 more than 45% of the time.

In Figure 7b we have plotted the mean clarity scores for each of the 12 sentences considered in the survey, for avatar animations and videos respectively. Scores for avatar animations were consistently lower than for videos, as expected, but the difference between the two varied considerably across sentences. In particular, the figure shows that while all sentences received roughly the same mean score when signed by a human signer (mean 9.0, standard deviation 0.223), the variability among sentences was considerably larger when signed by the avatar (mean 6.4, standard deviation 1.1).

When we zoom in on the scores for avatar animations, we see that sentences 9 and 12 may be considered negative outliers, since their mean scores (3.7 and 5.0, respectively) were more than one standard deviation below the mean across all sentences. Glosses of these two sentences are given in (9) and (10), respectively:

(9) Sentence 9: PAST SEVEN DAY YOU ALREADY CORONA TEST

(10) Sentence 12: SORRY I FAIL YOU INTRAVENOUS DRIP I COLLEAGUE CALL

Sentence 9 starts with the phrase PAST SEVEN DAY, which we already saw was very difficult for participants to recognize (see Table 4). Similarly, sentence 12 also contains three signs which were difficult to recognize when signed by the avatar, SORRY, FAIL, and INTRAVENOUS DRIP. Plausibly, these individual signs contributed to the low clarity scores for these sentences.

If sentences 9 and 12 are disregarded, the mean clarity score for videos is still 9.0 while the mean score for avatar animations rises to 6.8. A possible interpretation of the clarity score data, then, is that there is a 'root difference' between videos and avatar animations of about 2.2 points (the difference between the two means if the two outlier sentences are disregarded), and that this difference becomes larger when a sentence contains individual signs or phrases which are particularly difficult to recognize when signed by the avatar in the current implementation of the system. We hypothesize that the latter effect may in principle be reduced by improving the way in which these individual signs and phrases are rendered by the avatar. On the other hand, we expect that the 'root difference' between videos and avatar animations will be more persistent. Closing this gap will not be a matter of 'quick fixes' but will require more fundamental improvements of the underlying avatar technology (e.g. by making use of motion capture instead of scripted animation, or a combination of the two).

4.3.3 Guided versus unguided

We now turn to results pertaining to our third research question, comparing results of the GUIDED group with those of the UNGUIDED group.

Comprehension rates for individual signs We define the comprehension rate of an individual sign as the number of times that the sign was correctly recognized (code 4) divided by the total number of responses for that sign (codes 0-4). So the comprehension rate is a number between 0 and 1, reflecting the proportion of cases in which the sign was correctly recognized. Table 5a shows the mean comprehension rate of all individual signs, split by participant group (GUIDED vs UNGUIDED) and signer (**Av**atar vs **Hu**man). We see that for the GUIDED group, the mean comprehension rate

	GUIDED		UNGUIDED	
	Av	**Hu**	**Av**	**Hu**
Mean	0.82	0.87	0.90	0.96
St. dev.	0.38	0.34	0.30	0.20

(a) Individual signs

	GUIDED		UNGUIDED	
	Av	**Hu**	**Av**	**Hu**
Mean	0.68	0.95	0.80	0.99
St. dev.	0.47	0.21	0.40	0.11

(b) Sentences

Table 5: Effect of guidance on comprehension rates for individual signs and sentences

was only 0.05 lower for the avatar (mean 0.82, standard deviation 0.38) than for the human signer (mean 0.87, standard deviation 0.34). For the UNGUIDED group, the mean comprehension rate for the avatar was again close to that of the human signer, with a difference of 0.06 (avatar mean 0.90, standard deviation 0.30; human signer mean 0.96, standard deviation 0.20).

If we compare the GUIDED and UNGUIDED group per signer, we see that the mean comprehension rate for the avatar was 0.08 higher in the UNGUIDED group than in the GUIDED group, while the mean comprehension rate for the human signer was 0.09 higher in the UNGUIDED group than in the GUIDED group.

Overall, then, results from the GUIDED and the UNGUIDED group were quite similar when it came to the mean comprehension rate of individual signs.

Comprehension rates for full sentences We define the comprehension rate of a sentence as the number of times that the sentence was interpreted as intended (code 4) divided by the total number of responses elicited for that sentence (codes 0-4). So, just like in the case of individual signs, the comprehension rate for a sentence is a number between 0 and 1, reflecting the proportion of cases in which the sentence was correctly interpreted.

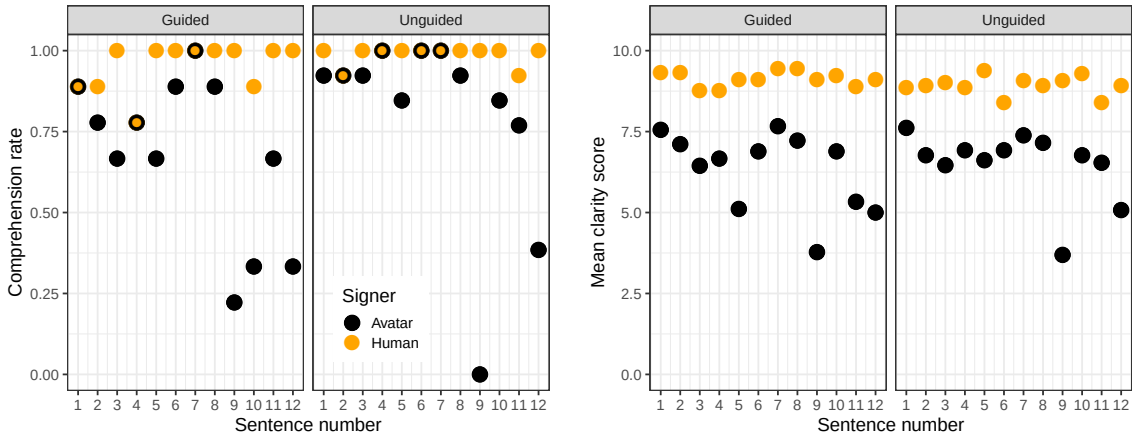
Table 5b shows the mean comprehension rate of all sentences, split by participant group (GUIDED vs UNGUIDED) and signer (**Av**atar vs **Hu**man). We see that for the GUIDED group, the mean comprehension rate was 0.27 lower for the avatar (mean 0.68, standard deviation 0.47) than for the human signer (mean 0.95, standard deviation 0.21). For the UNGUIDED group, the mean comprehension rate for the avatar was 0.19 lower for the avatar (mean 0.80, standard deviation 0.40) than for the human signer (mean 0.99, standard deviation 0.11).

If we compare the GUIDED and UNGUIDED group per signer, we see that the mean comprehension rate for the avatar was 0.12 higher in the UNGUIDED group than in the GUIDED group, while the mean comprehension rate for the human signer was 0.04 higher in the UNGUIDED group than in the GUIDED group.

Overall, then, results from the GUIDED and the UNGUIDED group were quite different when it came to the mean comprehension rate of sentences signed by the *avatar*. In this case, rates were substantially higher in the UNGUIDED group than in the GUIDED group. On the other hand, the comprehension rates of sentences signed by a *human* were more similar across the two groups, closer to what we observed for individual signs.

We further observe that, both in the GUIDED and in the UNGUIDED group, the standard deviation of the comprehension rates for sentences signed by the avatar was much higher than that of the comprehension rates for sentences signed by a human.

To obtain a better understanding of this larger variance in comprehension rates, Figure 8a plots the mean comprehension rates of all sentences, signed by the avatar and a human signer,



(a) Sentence comprehension rates

(b) Clarity scores

Figure 8: Effect of guidance on sentence comprehension rates and clarity scores assigned to sentences

respectively. The left pane (for the GUIDED group) shows that 8 sentences had a relatively high comprehension rate, while four sentences had a relatively low comprehension rate. Similarly, in the right pane (for the UNGUIDED group) we see that 9 sentences had a high comprehension rate, and three a low one. Overall, there is great similarity between the GUIDED and the UNGUIDED group as to which sentences received high rates and which ones received low rates. The only exception is sentence 10, which had a low rate in the GUIDED group but a high rate in the UNGUIDED group.

Clarity scores Figure 8b compares the GUIDED and the UNGUIDED group with respect to clarity scores. We see that the left pane (for the GUIDED group) is overall very similar to the right pane (for the GUIDED group), and both panes are similar to the plot in Figure 7b above, which displayed clarity scores for all sentences without making a distinction between the GUIDED and the UNGUIDED group. The only salient difference between the left and the right pane in Figure 8b pertains to sentences 5 and 11. These sentences received lower clarity scores in the GUIDED group than in the UNGUIDED group.

The overall impression that arises from comparing the GUIDED and the UNGUIDED group w.r.t. comprehension rates and clarity scores is that the two yielded largely similar results. The only case in which we observed a substantial difference between the two groups was in the comprehension rates for sentences signed by the avatar. In Section 4.3.4 below we will analyse in which cases the effect of guidance on comprehension rates and clarity scores was statistically significant.

4.3.4 Statistical analysis

We now present statistical models that investigate whether comprehension rates and clarity scores were significantly affected by *guidance* (GUIDED vs. UNGUIDED) and the type of *signer* (avatar vs. human). Three statistical models are considered: one for the comprehension of individual signs, one for the comprehension of sentences and one for clarity scores. The models are built in R using

the lme4 package (Bates et al., 2014).

Comprehension of individual signs We concentrate on the binary distinction between correct responses on the one hand (code 4) and incorrect or missing responses on the other hand (code 0-3). We applied a generalized linear mixed effects logistic regression, which models a binary outcome as a combination of predictors. These predictors are labeled either as fixed effects or as random effects (for an introduction into mixed effects models in linguistics, see Winter 2019; for a more advanced presentation, see Gelman and Hill 2007). In our case, the outcome is the response (0 = incorrect or missing response, 1 = correct response). Random effects are elements that would vary from one experiment to the next (if the same question was investigated) and which the experimenters do not exert full control over, either because this is impossible or because it is not of interest. In our case, participants and the signs that were used in the experiment are random effects. Fixed effects, on the other hand, are experimental manipulations that are of main interest and that would not vary from one experiment to the next. In our case, the fixed effects are *guidance* (with GUIDED and UNGUIDED as values, coded in the model using sum contrast coding as 0.5 and -0.5, respectively) and *signer* (with human and avatar as values, coded in the model using sum contrast coding as 0.5 and -0.5, respectively). We also include the interaction of *guidance* and *signer*, which reveals whether the difference due to one fixed effect (say, *guidance*) differs across different values of the other fixed effect (i.e., *signer*). Following common practice, we use the mixed effects model with the maximal random-effect structure that converges (Barr et al., 2013), which in our case was a model with a random intercept and random *signer* slopes (but no random *guidance* slopes) for participants and signs.

The obtained model identified both *signer* and *guidance* (but not their interaction) as significant predictors. The effect of *signer* was positive ($\beta = 4.1, z = 2.5, p = 0.013$), which means that the comprehension rate of signs was significantly higher when performed by the human signer than when performed by the avatar. The effect of *guidance* was negative ($\beta = -1.3, z = -5.3, p < 0.0015$), which means that comprehension rates for individual signs were significantly higher in the UNGUIDED group than in the GUIDED group. Finally, the fact that the interaction between *signer* and *guidance* was not a significant predictor means that we have no evidence that the effect *signer* differed across the GUIDED and the UNGUIDED group.

Sentence comprehension We again assumed a binary outcome variable, only distinguishing between cases in which a sentence was interpreted as intended (outcome = 1) and cases in which is wasn't (outcome = 0). More fine-grained distinctions such as that between missing and incorrect interpretations were disregarded. Just as in the case of individual sign comprehension, we applied mixed effects logistic regression. This time, the random effects were participants and sentences. As before, the fixed effects were *guidance* (with GUIDED and UNGUIDED as values, coded in the model using sum contrast coding as 0.5 and -0.5, respectively), *signer* (with human and avatar as values, coded in the model using sum contrast coding as 0.5 and -0.5, respectively), and the interaction between *guidance* and *signer*. The model with the most comprehensive random-effect structure that converged included a random intercept for participants and sentences and random *signer* slopes for sentences (but not for participants).

The model revealed a significant positive effect of *signer* ($\beta = 2.6, z = 2.9, p = 0.004$), which means that the comprehension rate of sentences was significantly higher when signed by a human than when signer by the avatar. This matches the positive effect of *signer* on the comprehension of individual signs that we found above.

Clarity scores Clarity was judged on a discrete scale from 0 to 10. We applied ordered probit mixed effects regression in this case, which models a discrete and bounded set of outcomes as resulting from a combination of fixed and random effects (see Kruschke 2011 for an introduction to ordered probit mixed effects regression models). Just like the model for sentence comprehension, this model included participants and sentences as random effects and *guidance* and *signer*, as well as the interaction between *guidance* and *signer*, as fixed effects. The model with the most comprehensive random-effect structure that converged included a random intercept and random *signer* slopes for participants and sentences.

The model identified only one significant effect. Namely, *signer* had a positive effect ($\beta = 4.3, z = 7.6, p < 0.001$), which means that clarity scores for sentences performed by a human signer were significantly higher than those for sentences performed by the avatar. This matches the positive effect of *signer* on the comprehension rates for sentences and individual signs that we found above.

4.3.5 Attitudes towards signing avatars

We now turn to results obtained in Part 4 of the survey. This part did not concern the comprehensibility or clarity of the avatar in our prototype system, but rather probed participants' attitudes towards signing avatars in general, and also asked them for feedback on the setup of our survey.

Participants' general attitude towards signing avatars was generally positive: 86.4% found that signing avatars should be further investigated and could potentially be very useful when further developed (see Table 6). Many participants noted explicitly that the technology in it's current state is not advanced enough yet to be deployed in real-life settings. Moreover, multiple participants explicitly commented that the looks of the avatar in our current prototype need to be improved. It was perceived as rather stiff, not very friendly, sometimes even scary. Such comments are reminiscent of the well-known 'uncanny valley' effect.

Those participants who believed that signing avatars should *not* be investigated further (13.6%) felt that it would be impossible for an avatar to ever display human-like facial expressions, and believed that avatars might take away jobs from human sign language interpreters and teachers.

Participants were also asked to reflect on possible use cases for signing avatars. A few example use cases were given in the survey (see Table 6 and Appendix A.4.1). Most participants indicated that signing avatars could be useful in public places such as train stations and airports, to relay travel information to passengers (77.2% in favour, 22.7% against). Opinions were divided about the use of signing avatars in medical settings (50.0% in favour, 50.0% against) or as part of tools to support people in learning sign language (31.8% in favour, 68.2% against). Concerning both these use cases, some participants were very enthusiastic, but others were strongly opposed. Participants on both sides of the spectrum indicated that the technology would need to be improved significantly before being used in these settings.

In addition to the example use cases listed in the survey, participants suggested other possible use cases as well. Multiple participants indicated that a signing avatar could be useful in waiting rooms, for standardised governmental processes and public services (e.g. renewing a passport), and in supermarkets.

4.3.6 Feedback on the setup of the survey

All participants indicated that the questions in the survey were clearly formulated: mean = 9.22 (standard deviation = 1.09) in the GUIDED group and mean = 8.38 (standard deviation = 1.04) in

		In favour	Against
Attitude	Should be investigated further?	86.4	13.6
Use cases	Travel information	77.2	22.7
	Medical settings	50.0	50.0
	Support for learning	31.8	68.2

Table 6: Participants' attitude towards signing avatars, and possible use cases

the UNGUIDED group) and that it was easy to provide answers (mean of 8.78, standard deviation of 1.30 in the GUIDED group and mean of 8.23, standard deviation of 1.01 in the UNGUIDED group).

Moreover, participants from both groups indicated that they felt taken seriously, although GUIDED participants commented on this more often than UNGUIDED ones. The fact that all instructions and questions in the survey were given in two formats, Dutch text and NGT videos, the fact that these videos featured a deaf signer, and the fact that the transcription of responses involved a feedback loop in the case of the GUIDED group, as described in Section 4.2.2, were all mentioned as contributing to this feeling of being taken seriously.

Some participants from both groups regretted that they were not explicitly asked for suggestions on how to improve the signing quality of the avatar. Systematically collecting such input was intentionally left outside the scope of the survey, to keep sessions manageable in terms of time and cognitive effort. Regardless, some participants from the GUIDED group did provide suggestions, and the experimenters made note of these.

Opinions on the added value of being able to respond in NGT, as opposed to entering responses textually, were divided. On the one hand, 92% of participants in the UNGUIDED group, who had to enter their responses textually, indicated that it would *not* have been easier to respond in NGT. On the other hand, while this question was not explicitly posed to participants in the GUIDED group, since they did respond in NGT, 50% of these participants spontaneously mentioned that they appreciated being able to use NGT throughout the survey and not having to enter their responses textually. We note that these findings may well be due, at least in part, to the fact that the participants themselves determined whether they would be part of the GUIDED or the UNGUIDED group.

4.4 Discussion

We now discuss the results of the survey in light of our three main research questions, reflect on the limitations of the conclusions that can be drawn from these results, and suggest some avenues for future work.

RQ1: Comprehension of individual signs The comprehension rate of individual sings performed by the avatar (86.93%) was lower than the comprehension rate of individual signs performed by a human signer (91.34%), but both rates are quite high and the difference between them is rather small. A closer look at the most common mistakes revealed, among other things, that the *mouthings* produced by the avatar can be particularly confusing. For instance, SOMETIMES was misinterpreted 54.5% of the time as MAYBE, which has the same manual component but different mouthing, and

similarly, RESULT was misinterpreted 22.7% of the time as PASSED. Such misinterpretations did not arise when the signs were performed by a human signer.

The JAsigning avatar engine currently offers limited possibilities to produce natural-looking mouthings. More specifically, the engine currently allows for a specification of mouthings in SAMPA notation (Speech Assessment Methods Phonetic Alphabet). This is a phonetic notation system: each SAMPA symbol corresponds to a particular phoneme. There is, however, no one-to-one mapping between phonemes and mouth shapes. For instance, the 's' in 'sun' and the 's' in 'silver' involve the same phoneme but different mouth shapes because the next vowel is anticipated. This makes it difficult to generate correct mouthings in JAsigning. In future work, it would therefore be advisable to reconsider the way in which mouth shapes are handled in the engine. This line of future work may take inspiration from lipsync algorithms for game characters (e.g., Edwards et al. 2016).

RQ2: Sentence comprehension and clarity Sentences signed by the avatar had a comprehension rate of 74.62% and a mean clarity score of 6.4, while ones signed by a human had a comprehension rate of 97.35% and a mean clarity score of 9.0. So here we saw a larger contrast between the avatar and the human signer than in the case of individual signs.

Taking a closer look at the 12 sentences that were used in the survey, we found that 2 of these received particularly low scores when signed by the avatar. These sentences were ones which contained several individual signs with low comprehension rates. Improving the way in which these particular signs are rendered by the avatar may well improve the scores of the sentences that contained them as well.

However, we noted that even if these 2 negative outlier sentences were to be disregarded, there is still a substantial difference between the comprehension rates and clarity scores for the avatar and those for the human signer. To close this gap, it will not suffice to improve the rendering of some individual signs. Rather, more fundamental improvements of the underlying avatar technology will be necessary. Based on the feedback provided by participants during the survey, we note that sentence prosody (the relative speed and intensity of the signs in the sentence) and the transitions between signs are important elements that strongly influence comprehensibility. The JASigning avatar engine and the SiGML formalism that it makes use of currently offer limited possibilities to control prosody and transitions. To make the engine suitable for practical applications, these functionalities need to be extensively developed in future work. An alternative would be to explore an approach that makes use of motion capture instead of scripted animation (see, e.g., Gibet, 2018), or a combination of the two.

RQ3: Effect of guidance In previous work, the evaluation of signing avatars generally involved 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). However, the COVID-19 pandemic made it necessary for us to turn to online procedures, and it is to be expected that in the future researchers may sometimes want to employ online procedures as well. This new experimental setting raises methodological issues. In particular, one basic design choice that needs to be made concerns the online presence of the experimenters while the participants take the survey. Outside the domain of sign language technology, it is most common in online quantitative surveys for the experimenters *not* to be present. For the specific purpose of evaluating signing avatars, however, this has a possible disadvantage, namely that participants have to enter their responses textually rather

than in sign language. This may be dispreferred, at least for some participants. To circumvent this potential disadvantage, we offered our participants a choice between a GUIDED and an UNGUIDED version of the survey. A comparison between the results from the GUIDED and the UNGUIDED group, as well as the feedback that participants from both groups provided on the setup of the survey, may inform the design of future online evaluation studies.

The main difference between the GUIDED and the UNGUIDED group was that comprehension rates, both for individual signs and for sentences, were generally *lower* in the GUIDED group. This was an unexpected result for us. If anything, we had expected comprehension rates to be higher in the GUIDED group. There is, however, a plausible explanation for why there was in fact a difference in the opposite direction. Namely, it may be that the presence of the experimenters caused a certain amount of social pressure for participants in the GUIDED group. For instance, they may have felt that it would be a burden to ask the experimenters to replay a video, or they may have felt pressure to understand sentences and signs on the first try. They may even have experienced the experiment partly as a memory task rather than a pure comprehension task. Participants from the UNGUIDED group presumably did not experience any such pressures, and may have felt more freedom to replay videos as often as needed. This may be one of the reasons that comprehension rates were higher in the UNGUIDED group.

As for the feedback we received from both groups on the setup of the survey, 92% of participants from the UNGUIDED group indicated that it would *not* have been easier to respond in NGT instead of entering their responses textually. On the other hand, while this question was not explicitly posed to participants in the GUIDED group, since they did respond in NGT, 50% of them spontaneously mentioned that they appreciated being able to use NGT throughout the survey and not having to enter their responses textually.

Overall, then, it is not the case that a GUIDED online procedure is to be strictly preferred over an UNGUIDED procedure for the evaluation of signing avatars, nor vice versa. Both methods have advantages and disadvantages, and which format works best differs from one participant to another. For future work, we can therefore only recommend that, whenever a choice needs to be made between a GUIDED and an UNGUIDED setup, the potential advantages and disadvantages of both options are carefully weighed.

Limitations There are various factors that limit the generalisability of the results of our survey. First, the design of our survey allowed for two kinds of learning effect to arise. On the one hand, each participant first saw twelve sentences signed by the avatar and then the same twelve sentences signed by a human signer. This may in part explain why the human signer received higher comprehension rates and clarity scores than the avatar. On the other hand, some individual signs appeared in more than one sentence and were therefore seen more often than other signs. This may have positively affected their comprehension rate.

Second, the differences we found between the GUIDED and the UNGUIDED group may in part be due to the fact that participants chose themselves whether to take the GUIDED or the UNGUIDED version of the survey. For instance, this may in part explain why 92% of participants from the UNGUIDED group indicated that it would not have been easier to respond in NGT instead of entering their responses textually.

Finally, some more general limitations apply: our survey involved only a small number of sentences and signs, a rather small participant pool, and was conducted in a controlled environment rather than in a real-life setting. When interpreting our results, these factors should be kept in mind, and future work should investigate how well the results generalise.

5 Conclusion

We have investigated the potential of automated text-to-sign translation to address the challenges that the COVID-19 pandemic implies for the communication between healthcare professionals and deaf patients. We motivated a modular approach to automated text-to-sign translation, and implemented a first prototype system. We conducted a survey among potential end-users to evaluate the comprehensibility and clarity of the avatar. Moreover, we investigated whether the possibility to interact with the experimenters during the survey and to provide responses in NGT rather than having to enter them textually affected the results when conducting a survey of this sort in an online environment. We have discussed various prospects and limitations of the prototype system we built and of the results of our survey.

For the approach taken here to become viable in practice, the JASigning avatar engine needs to be substantially further developed. At the level of individual signs, the engine should allow for more subtle body movements and facial expressions, and the system for encoding mouth shapes should be revised. At the level of sentences, more control is needed to adapt the relative speed and intensity of the different signs within a sentence (prosody) and to make transitions between signs more natural and smooth. An alternative is to explore an approach based on motion capture instead of scripted animation, or a combination of both.

Finally, we believe that future projects will strongly benefit from a more inclusive and more iterative design process, involving a multi-disciplinary team with a strong representation of deaf researchers and domain experts. The design and implementation phase of the present project was carried out under great time pressure, given the urgency of the issue we aimed to address, and with limited resources. In future work, there should be several design iterations, the design team should include deaf specialists on sign language and Deaf Studies from the start, and focus groups should be organised to receive input from a larger group of potential end-users, aiming for maximal diversity in terms of age, region, and level of education.

Acknowledgements We are grateful to Tashi Bradford, Richard Cokart, Bastien David, John Glauert, Lisa Hinderks, Richard Kennaway, Lisa van der Mark, Marta Morgado, Joni Oyserman, Marijke Scheffener, Anique Schüller, and Roos Wattel for their help at various stages of this project.

Funding We gratefully acknowledge financial support from the Netherlands Organisation for Innovation in Healthcare (ZonMW), the Netherlands Organisation for Scientific Research (NWO), and the European Research Council (grant number 680220).

Declarations

Competing interests All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Consent to participate Informed consent was obtained from all subjects for being included in the study.

References

Baker, A., van den Bogaerde, B., Pfau, R., and Schermer, T. (2016). *The linguistics of sign languages: An introduction*. John Benjamins Publishing Company.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1.

Coerts, J. (1992). *Nonmanual grammatical markers: an analysis of interrogatives, negations and topicalisations in Sign Language of the Netherlands*. PhD thesis, University of Amsterdam.

Courty, N. and Gibet, S. (2010). Why is the creation of a virtual signer challenging computer animation? In *International Conference on Motion in Games*, pages 290–300. Springer.

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.

de Meulder, M. and Haualand, H. (2021). Sign language interpreting services: A quick fix for inclusion? *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association*, 16(1):19–40.

de Vos, C., van der Kooij, E., and Crasborn, O. (2009). Mixed signals: Combining linguistic and affective functions of eyebrows in questions in Sign Language of the Netherlands. *Language and Speech*, 52(2-3):315–339.

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.

Edwards, P., Landreth, C., Fiume, E., and Singh, K. (2016). JALI: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics*, 35(4):1–11.

Elliott, R., Glauert, J., Jennings, V., and Kennaway, R. (2004). An overview of the SiGML notation and SiGML signing software system. In *Workshop on the Representation and Processing of Sign Languages at the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 98–104.

Esselink, L. (2020). Lexical resources for sign language synthesis: The translation of Dutch to Sign Language of the Netherlands. Bachelor's thesis. University of Amsterdam, https://scripties.uba.uva.nl/search?id=715792.

Fellinger, J., Holzinger, D., and Pollard, R. (2012). Mental health of deaf people. *The Lancet*, 379(9820):1037–1044.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press.

Gibet, S. (2018). Building French Sign Language motion capture corpora for signing avatars. In *Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, LREC 2018*.

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.

Glauert, J. and Elliott, R. (2011). Extending the SiGML notation–a progress report. In *Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT)*.

Grote, H. and Izagaren, F. (2020). COVID-19: The communication needs of D/deaf healthcare workers and patients are being forgotten. *British Medical Journal*, 369.

Hanke, T. (2004). HamNoSys-representing sign language data in language resources and language processing contexts. In *LREC*, volume 4, pages 1–6.

Hou, L. and de Vos, C. (2022). Classifications and typologies: Labeling sign languages and signing communities. *Journal of Sociolinguistics*, 26(1):118–125.

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

Jennings, V., Elliott, R., Kennaway, R., and Glauert, J. (2010). Requirements for a signing avatar. In *Workshop on Corpora and Sign Language Technologies at the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 33–136.

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.

Klomp, U. (2021). *A descriptive grammar of Sign Language of the Netherlands*. PhD thesis, University of Amsterdam.

Kruschke, J. K. (2011). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press/Elsevier.

Kusters, A. and Lucas, C. (2022). Emergence and evolutions: Introducing sign language sociolinguistics. *Journal of Sociolinguistics*, 26(1):84–98.

McKee, M., Moran, C., and Zazove, P. (2020). Overcoming additional barriers to care for deaf and hard of hearing patients during covid-19. *JAMA Otolaryngology–Head & Neck Surgery*, 146(9):781–782.

McKee, M. M., Paasche-Orlow, M. K., Winters, P. C., Fiscella, K., Zazove, P., Sen, A., and Pearson, T. (2015). Assessing health literacy in deaf American Sign Language users. *Journal of health communication*, 20(sup2):92–100.

Middleton, A., Niruban, A., Girling, G., and Myint, P. K. (2010). Communicating in a healthcare setting with people who have hearing loss. *Bmj*, 341.

Napier, J. and Kidd, M. R. (2013). English literacy as a barrier to health care information for deaf people who use auslan. *Australian family physician*, 42(12):896–899.

Prillwitz, S., Leven, R., Zienert, H., Hanke, T., and Henning, J. (1989). HamNoSys Version 2: Hamburg Notation System for Sign Languages — An Introductory Guide. *International Studies on Sign Language and the Communication of the Deaf*, 5.

Prins, M. and Janssen, J. B. (2014). Automated sign language. TNO technical report.

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:https://doi.org/10.31234/osf.io/g2wuc.

Rayner, M., Bouillon, P., Ebling, S., Gerlach, J., Strasly, I., and Tsourakis, N. (2016). An open web platform for rule-based speech-to-sign translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 162–168, Berlin, Germany. Association for Computational Linguistics.

Roelofsen, F., Esselink, L., Mende-Gillings, S., De Meulder, M., Sijm, N., and Smeijers, A. (2021a). Online Evaluation of Text-to-sign Translation by Deaf End Users: Some Methodological Recommendations (short paper). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 82–87.

Roelofsen, F., Esselink, L., Mende-Gillings, S., and Smeijers, A. (2021b). Sign language translation in a healthcare setting. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 110–124.

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*.

Smeijers, A. S., Ens-Dokkum, M. H., van den Bogaerde, B., and Oudesluys-Murphy, A. M. (2011). Clinical practice: The approach to the deaf or hard-of-hearing paediatric patient. *European journal of pediatrics*, 170(11):1359.

Smeijers, A. S. and Roelofsen, F. (2021). Communicatiebehoeften en ervaringen van dove patiënten in Nederland tijdens de COVID-19 pandemie (communication needs and experiences of deaf patients in the netherlands during the covid-19 pandemic). Dutch report available through https://zorgbeter.info/, with per-section summaries in NGT.

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.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.

Wolfe, R., Cook, P., McDonald, J. C., and Schnepp, J. (2011). Linguistics as structure in computer animation: Toward a more effective synthesis of brow motion in American Sign Language. *Sign Language & Linguistics*, 14(1):179–199.

Zwitserlood, I. (2005). Synthetic signing. In *The World of Content Creation, Management, and Delivery (IBC 2005)*, page 352–357.

A Survey

A.1 Part 1 – General questions

1. What is your age?
   A. 18-30
   B. 31-40
   C. 41-50
   D. 51+
2. What is your gender?
   A. Male
   B. Female
3. Which region are you from?
   A. Central (Utrecht, Flevoland)
   B. North (Groningen, Friesland, Drenthe)
   C. East (Gelderland, Overijssel)
   D. South (Brabant, Zeeland, Limburg)
   E. West (Noord-Holland, Zuid-Holland)
4. What is your mother tongue?
   A. NGT
   B. Dutch
   C. Both
   D. Other, namely ...
5. How often do you use NGT?
   A. Daily
   B. Regularly (a few days a week)
   C. Occasionally (a few days a month)
   D. Rarely (a few days a year)
6. How often do you make use of an interpreter?
   A. Daily

B. Regularly (a few days a week)

C. Occasionally (a few days a month)

D. Rarely (a few days a year)

7. How often do you communicate with people that do not use sign language, without making use of an interpreter NGT?

   A. Daily

   B. Regularly (a few days a week)

   C. Occasionally (a few days a month)

   D. Rarely (a few days a year)

8. How do you communicate with people that do not use sign language when an interpreter cannot be present? (multiple options possible)

   A. Lipreading

   B. Speech recognition through the phone

   C. Writing on paper/phone

   D. Signing

   E. Using voice

   F. Other, namely...

A.2 Part 2 – Avatar comprehension and clarity

The format of the questions in Part 2 is shown in Figure 4.

1. Did you sleep well?
   YOU GOOD SLEEP PALMS-UP
   https://youtu.be/WWJlTBBofZs

2. Do you use any medications?
   YOU MEDICINE USE PALMS-UP
   https://youtu.be/AybaxFnZuPk

3. What are you allergic to?
   YOU ALLERGIC FOR WHAT PALMS-UP
   https://youtu.be/Kutcmt7FWjo

4. Please stay in bed.
   YOU PLEASE BED STAY
   https://youtu.be/WNQi7QYYi2Q

5. I will come back later.
   I LATER BACK
   https://youtu.be/Qwn9e9TGOrQ

6. A colleague will come by soon to draw blood.
   SOON COLLEAGUE COME BLOOD DRAW
   https://youtu.be/hnKHRxexQgo

7. Who is your general practicioner?
   YOU GENERAL-PRACTICIONER WHO INDEX
   https://youtu.be/3DKEm4NzZzg

8. Do you have hearing aids or a cochlear implant?
   HEARING-AID OR COCHLEAR-IMPLANT HAVE YOU INDEX PALMS-UP
   https://youtu.be/KKTbJZRceuA

9. Have you had a Corona test in the past seven days?
   PAST SEVEN DAY YOU ALREADY CORONA TEST INDEX PALMS-UP
   https://youtu.be/E0N119HEUUs

10. Your Corona test results are negative.
    YOU CORONA TEST RESULTS NEGATIVE
    https://youtu.be/uchqMCv5eL0

11. Sometimes the test is wrong, therefore we have to do more research.
    SOMETIMES TEST WRONG THEREFORE WE MORE RESEARCH HAVE-TO
    https://youtu.be/vIbkt3Mp4xw

12. Sorry, I'm failing to insert the intravenous drip. I'll call a colleague.
    SORRY INTRAVENOUS-DRIP FAIL I COLLEAGUE CALL
    https://youtu.be/WvARK-HWDOU

A.3 Part 3 – Human signer comprehension and clarity

The format of the questions in Part 3 is shown in Figure 4.

1. Did you sleep well?
   YOU GOOD SLEEP INDEX
   https://youtu.be/CcY_u6Wh784

2. Do you use any medications?
   YOU USE MEDICINE INDEX
   https://youtu.be/C8GfYQoxRek

3. What are you allergic to?
   YOU ALLERGIC FOR WHAT PALMS-UP
   https://youtu.be/BssLQ9qva1g

4. Please stay in bed.
   YOU BED STAY
   https://youtu.be/biZrDOESgOc

5. I will come back later.
   I LATER BACK
   https://youtu.be/am9qBU9Ilf8

6. A colleague will come by to draw blood.
   COLLEAGUE COME BLOOD DRAW
   https://youtu.be/Z6G5HFg1ZbY

7. Who is your general practicioner?
   YOU GENERAL-PRACTICIONER INDEX WHO INDEX PALMS-UP
   https://youtu.be/fBKTJ29SLBU

8. Do you have hearing aids or a cochlear implant?
   HEARING-AID OR COCHLEAR-IMPLANT HAVE YOU
   https://youtu.be/pSseoqw2HdE

9. Have you had a Corona test in the past seven days?
   PAST SEVEN DAY YOU ALREADY CORONA TEST PALMS-UP
   https://youtu.be/XuiLZ9nRlO0

10. Your Corona test results are negative.
    YOU CORONA TEST RESULTS YOU NEGATIVE
    https://youtu.be/nlE9Ci_GJ24

11. Sometimes the test is wrong, therefore we have to do more research.
    SOMETIMES TEST INDEX WRONG WE MORE RESEARCH HAVE-TO
    https://youtu.be/N3I9O7FrnS0

12. Sorry, I'm failing to insert the intravenous drip. I'm calling a colleague.
    SORRY I FAIL YOU INTRAVENOUS DRIP I COLLEAGUE CALL
    https://youtu.be/80x_N9z85P8

A.4 Part 4 – Final questions

A.4.1 Attitude towards signing avatars

1. There has not been much research on avatar technology for translating text to sign language. Do you think this research should be continued and this technology should be developed further?

   A. Yes, because ...

   B. No, because ...

2. In which situations do you think that avatar technology for translating text to sign language can help? (multiple answers possible)

   A. For translating travel information in trains and on train stations

   B. For translating travel information in airplanes and at airports

   C. As support for people who want to learn sign language

   D. In hospitals during the COVID-19 crisis

   E. Other situations, namely ...

   F. I think that this technology is not helpful in any situation

A.4.2 Feedback on methodology

1. Were the questions in this study clearly formulated? (scale of 0-10)
2. Were the questions in this study easy to answer? (scale of 0-10)
3. Would it have been easier to answer questions in NGT rather than in Dutch text? (only for the UNGUIDED group)
   A. Yes, because ...
   B. No, because ...
4. Which aspects of the setup of the survey were pleasant?
5. Which aspects of the setup of the survey could be improved?

B User interface examples

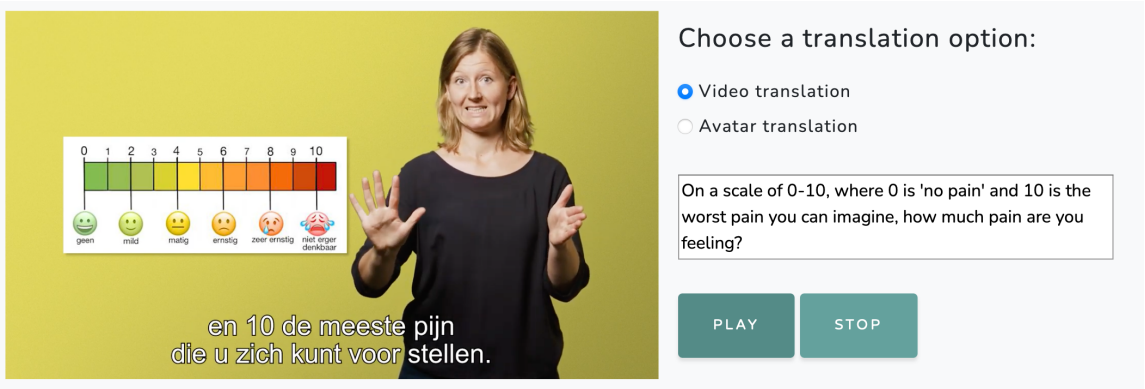
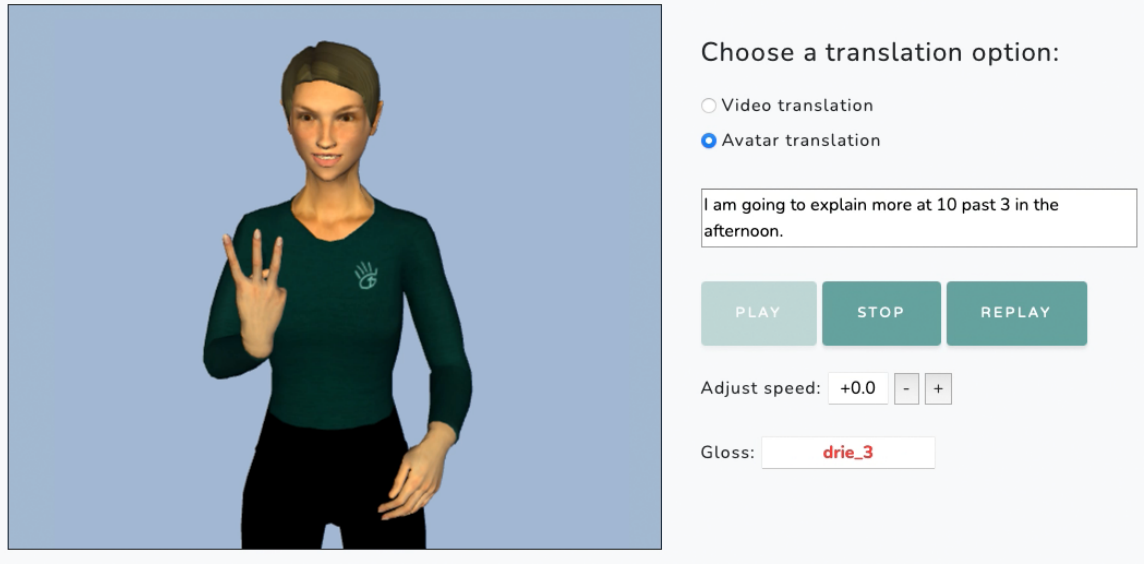


Figure 9: Example of a video translation of a complex question. The question is long and supported by an image

(a) Filling in values for variables



(b) Avatar signs sentence

Figure 10: If a sentence contains variables, the user is given options to insert the values of these variables (e.g. a clock, a number between 1-59). Once the user submits the values for the variables, the sentence on the interface is changed to reflect the chosen values, after which the avatar can sign the sentence