



D1.3 REPORT ON INTERIM EVALUATION STUDY

Revision: v.1.0

Work package	WP 1
Task	1.3
Due date	31/12/2022
Submission date	23/12/2022
Deliverable lead	EUD
Version	1.0
Authors	PICRON Frankie, VAN LANDUYT Davy, OMARDEEN Rehana (EUD)
Reviewers	HENAUULT-TESSIER Mélanie (INT), BRAFFORT Annelies (CNRS)
Abstract	This Deliverable is the report which presents the results from the first phase of evaluation studies in the EASIER project. The end-user evaluations were carried out by WP1 partners in all 7 project countries, with both deaf and hearing evaluation groups. Components from the EASIER project (application, translation and avatar) were evaluated and the feedback collected in this Deliverable, which serves as a reference for the Consortium to base future work on.
Keywords	Evaluation, app, application, translation, avatar, participants, feedback



Grant Agreement No.: 101016982
Call: H2020-ICT-2020-2
Topic: ICT-57-2020
Type of action: RIA

DISCLAIMER

The information, documentation and figures available in this deliverable are written by the "Intelligent Automatic Sign Language Translation" (EASIER) project's consortium under EC grant agreement 101016982 and do not necessarily reflect the views of the European Commission.

The European Commission is not liable for any use that may be made of the information contained herein.

COPYRIGHT NOTICE

© 2021 - 2023 EASIER Consortium

Project co-funded by the European Commission in the H2020 Programme		
Nature of the deliverable:		R*
Dissemination Level		
PU	Public, fully open, e.g. web	X
CL	Classified, information as referred to in Commission Decision 2001/844/EC	
CO	Confidential to EASIER project and Commission Services	

* R: Document, report (excluding the periodic and final reports)

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

OTHER: Software, technical diagram, etc.

EXECUTIVE SUMMARY

This report presents the results of the interim end-user evaluation carried out as part of Work Package 1 of the EASIER project. The end-user evaluation cycles were designed and carried out by members of the Work Package 1 group, under Task 1.2 with the aim of ensuring that end-user feedback is collected and integrated into the design of the EASIER system.

The evaluation was carried out across 7 countries, with 14 groups of deaf and hearing users from the project's target sign language communities. Three components of the EASIER system were evaluated in their prototype form, the app, the translation models and the avatar. Participants interacted with these systems and engaged in a facilitator-led discussion group to provide in depth feedback and recommendations.

RESULTS

Participants were pleased to be invited to participate and produced a great deal of feedback for each component. They complemented the positive aspects of the components, and provided specific feedback and suggestions for issues.

There was a considerable amount of qualitative feedback on the application. Participants brought up issues with several aspects of the app, and provided specific feedback on solutions, demonstrating a clear vision of what they wanted the app to look like. Major issues concerned progression through the app, setting up translation parameters, visual choices in icons and images, and general useability. Participants also were able to identify language specific localization issues. Overall, the app feedback provides a clear checklist of issues for developers to work through to increase useability and tailor the app towards user needs.

The translation feedback was less clear, as the technology behind machine translation was not well understood by participants. Nevertheless, this was ameliorated in many groups by including participants with a linguistics background who gave detailed feedback on specific issues present when moving from sign to spoken language, particularly when relying on written glosses which do not capture critical information such as non manuals. Participants also discussed in detail the issues with discourse context when establishing reference in sign to convey verb tense or pronoun/pointing referents.

Participants were quite precise with feedback on the avatar, and were able to precisely pinpoint many sources of the 'unnatural' and 'robotic' qualities of Paula's singing. Non manuals arose as critical for intelligibility, and a core issue that needs improvement. Participants also highlighted the need for smoother transitions between signs, and overall work to create more varied changes in signing speed. Participants also identified several language specific errors with particular signs, which have been compiled in the deliverable.

Participants in each section provided useful comments for improving evaluation protocols that will be integrated into upcoming evaluation rounds.

REPORT ORGANISATION

The document is divided into 7 sections. A brief **Introduction** is followed by the Methodology section which provides details on procedure and participants. Then results from the individual components are presented in the sections **App**, **Translation** and **Avatar**. **Other Feedback**

contains additional issues that emerged from the discussion, and **Conclusions** offers some final remarks and takeaways.

TABLE OF CONTENTS

1 INTRODUCTION	11
2 METHODOLOGY	12
2.1 Investigation method	12
2.1.1 Participants.....	12
2.1.1.1 Facilitators	12
2.1.1.2 Participant recruitment and profile.....	12
2.1.2 Procedure	14
2.1.3 Data analysis	15
3 APP 16	
3.1 General feedback and recommendations	17
3.1.1 App navigation.....	17
3.1.2 Need for guidance	17
3.1.3 Translation screen and language selection.....	18
3.1.4 The use of flags	19
3.1.5 SL Video recording and minimum quality requirements.....	20
3.1.6 Audio recording	21
3.1.7 Text input.....	21
3.1.8 App settings and clarity of icons.....	22
3.1.9 The voice of machine translation.....	25
3.1.10 Dark mode and accessibility.....	26
3.1.11 Representation of diversity	26
3.1.12 Interaction flow	27
3.1.13 Archive, data security and privacy.....	28
3.1.14 Profiles, user IDs and logging in.....	29
3.1.15 Library of commonly used sentences.....	30
3.1.16 Choices in image design	30
3.1.17 Offline availability.....	31
3.2 Issue-specific feedback	32
3.2.1 Localization issues	32
3.2.2 Navigation issues	33
3.2.3 Design issues	34
3.2.4 Dark mode issues.....	37
3.2.5 Wording issues (across languages)	38
3.2.6 Technical issues	38
3.2.7 Other suggestions	39
3.3 Feedback on app evaluation protocols	40

4 TRANSLATION.....	41
4.1 Overall impression.....	41
4.2 Signed to spoken language	42
4.3 Spoken to signed language	45
4.4 Evaluation procedure/method.....	49
4.5 Questions	50
4.6 Recommendations.....	50
4.7 Conclusions.....	51
5 AVATAR	52
5.1 Questionnaire results.....	52
5.1.1 Single signs	53
5.1.2 Multi-sign utterances	53
5.2 Focus Group Discussion	55
5.2.1 Overall appearance and legibility	55
5.2.2 Non-manuals	56
5.2.3 Prosody	58
5.2.4 Manual sign formation	59
5.2.5 Methodological feedback.....	60
5.2.6 Recommendations.....	61
5.2.7 Other feedback	62
5.2.8 Conclusions	62
6 OTHER FEEDBACK.....	64
6.1 Use Cases.....	64
7 CONCLUSIONS.....	65

LIST OF FIGURES

FIGURE 1: DGS SETUP FOR DEAF (LEFT) AND HEARING (RIGHT) PARTICIPANTS. PARTICIPANTS USED DESKTOP COMPUTERS IN THE TESTING ROOM AND SAT IN A SEMICIRCLE FOR DISCUSSION WITH A LARGE SCREEN BEHIND THE FACILITATOR FOR SHOWING MATERIALS.....	15
FIGURE 2: LSF SETUP FOR DEAF (LEFT) AND HEARING (RIGHT) PARTICIPANTS. PARTICIPANTS BROUGHT USED LAPTOPS PROVIDED BY INTERPRETIS AND MOBILE PHONES AND SAT IN AN L (DEAF, LEFT) OR A U (HEARING, RIGHT) FORMATION, WITH A PROJECTOR PROJECTING ONTO THE WALL BEHIND THE FACILITATOR.	15
FIGURE 3: LOCATION OF THE BACK ARROW IN THE UPPER BAR	17
FIGURE 4: “BRITISH” IS SHOWN AS SELECTED LANGUAGE FOR AN INPUT THROUGH THE CAMERA.....	18
FIGURE 5: “BACK TO START MENU” BUTTON AFTER DOING A TRANSLATION	18
FIGURE 6: “DELETE”, “COPY”, “SAVE” AND “SHARE” BUTTONS AFTER DOING A TRANSLATION.....	19
FIGURE 7: SWISS FLAG ICON DISPLAYED AS INPUT LANGUAGE FOR A TEXT TRANSLATION	19
FIGURE 8: RECORD VIDEO SCREEN	21
FIGURE 9: AVATAR SETTINGS SCREEN.....	23
FIGURE 10: “USE FAVORITE SETTINGS” OPTION IN THE START SCREEN	23
FIGURE 11: VOICE SETTINGS SCREEN	24
FIGURE 12: INPUT MODALITY SELECTION	24
FIGURE 13: OPTIONS IN THE SETTINGS SCREEN	25
FIGURE 14: ACCENT OPTIONS FOR SPOKEN ENGLISH	25
FIGURE 15: GENDER OPTIONS IN THE SETTINGS	27
FIGURE 16: VISUALISED PROCESS OF AN INTERACTION FLOW BETWEEN INTERLOCUTORS WHILE USING THE CURRENT EASIER APP	28
FIGURE 17: ARCHIVE SCREEN	29
FIGURE 18: WORLD MAP IMAGE USED IN THE EASIER APP	30
FIGURE 19: CHOOSE LANGUAGE BUTTON IN THE EASIER APP	30
FIGURE 20: IMAGES USED IN THE EASIER APP.....	31
FIGURE 21: SETTINGS SCREEN.....	31
FIGURE 22: LANGUAGE OPTIONS IN ANOTHER LANGUAGE WHEN DUTCH IS SELECTED	32
FIGURE 23: “CLEAR TEXT” CONFIRMATION SCREEN IN ITALIAN	33
FIGURE 24: WORDS BEING BROKEN UP IN THE GERMAN VERSION OF THE APP.....	33
FIGURE 25: WORDS OVERLAPPING DESIGN IN THE ITALIAN VERSION OF THE APP	33
FIGURE 26: LANGUAGE SELECTION SCREEN WHICH SHOWS DSGS INSTEAD OF DGS	34
FIGURE 27: VIDEO RECORDING SCREEN DISPLAYING A MICROPHONE ICON.....	34
FIGURE 28: SIGN LANGUAGE OUTPUT SCREEN WITH THE AVATAR SIGNING OUTSIDE OF THE SCREEN	35

FIGURE 29: SETTINGS SCREEN.....	35
FIGURE 30: INPUT MODALITY SELECTION	36
FIGURE 31: START PAGE OF THE EASIER APP WITH THE DROP-DOWN MENU ALREADY OPEN	36
FIGURE 32: GREY TEXT IN THE EASIER LOGO IS BARELY VISIBLE IN DARK MODE (LEFT), BLACK STRIPE IN GERMAN FLAG SEEMS TO DISAPPEAR AND THE YELLOW TEXT ON A GREY BACKGROUND IS HARD TO READ (RIGHT)	37
FIGURE 33: CHOSEN LANGUAGE IS NOT VISIBLE IN THE DROP-DOWN MENU IN DARK MODE (LEFT), DROP-DOWN MENU APPEARS IN WHITE (RIGHT)	37
FIGURE 34: LANGUAGE SELECTION MENU IN DROP-DOWN MENU APPEARS IN WHITE WHILE IN DARK MODE.....	38
FIGURE 35: PARTICIPANT RATINGS ACROSS LANGUAGES FOR “DID THEY SIGN THE SAME?”	53
FIGURE 36: PARTICIPANT RATINGS ACROSS LANGUAGES FOR “HOW WELL DID PAULA SIGN?”	54
FIGURE 37: PARTICIPANT RATINGS ACROSS LANGUAGES FOR “DID YOU UNDERSTAND WHAT PAULA SIGNED?”	54
FIGURE 38: WRONG MOUTH GESTURE FOR DGS SIGN BALD1A.....	58
FIGURE 39: FALSE INTERMEDIARY SIGN BETWEEN UNDERSTAND (1) AND NOT (3) (LSF) ...	58
FIGURE 40: DISTORTED IMAGE OF HUMAN SIGNER (DSGS)	61

LIST OF TABLES

TABLE 1: FOCUS GROUPS SUMMARY TABLE	13
TABLE 2: DGS>DE TRANSLATION SENTENCES	42
TABLE 3: BSL>EN TRANSLATION SENTENCES	43
TABLE 4: DGS>DE TRANSLATION SENTENCES	45
TABLE 5: EN>BSL TRANSLATION SENTENCES	47
TABLE 6: SIGN SPECIFIC ERRORS	59
TABLE 7: AVATAR RECOMMENDATIONS.....	61

ABBREVIATIONS

BSL	British Sign Language
CODA	Child Of Deaf Adults
DGS	German Sign Language
DSGS	Swiss German Sign Language
EUD	European Union of the Deaf
GSL	Greek Sign Language
HT	Human translation
LSF	French Sign Language
LIS	Italian Sign Language
LSF	French Sign Language
MT	Machine translation
NGT	Dutch Sign Language
SL	Sign Language



1 INTRODUCTION

The interim user evaluation study falls under Task 1.3, **End user evaluation studies**, of Work Package 1. The interim user evaluation is the first of two user evaluations planned in the project lifespan, with the second final end-user evaluation to take place in M32 to M33. The aim of these evaluations is to collect end-user feedback from the project's target language communities on the intermediate stages of the EASIER components, specifically the mobile application, the translation models and the sign language avatar.

Two adjustments to the evaluation have been made from the original project proposal, as noted in Deliverable 1.2 **Report on performance metrics and user study preparations**.

First, the number of participants has been reduced from 30 to between 4-6 participants per language group. While it was initially planned to recruit a large sample of users for a remote evaluation framed by a questionnaire, the decision was made to instead focus on collecting qualitative feedback via focus groups conducted with smaller groups of participants. This shift of approach from a more quantitative to a more qualitative approach proved more suitable for an intermediate evaluation, as it gave the opportunity to collect highly nuanced and detailed feedback from users on 'beta' versions of the EASIER components. Despite being more time consuming and logistically challenging than the original plan, and involving fewer end users, the results clearly demonstrate the value of qualitative feedback at this stage of evaluation, as users provided highly nuanced feedback and recommendations for all components. This is perhaps most clearly observed for the avatar component, where there was also a quantitative questionnaire accompanying the discussion. Here, the questionnaire provided overall ratings from users, however the follow up discussions allowed users to explain in detail what criteria they used to rate the sentences.

Second, given the challenges in participant recruitment and carrying out the evaluations over the summer holiday, the deadline of the deliverable has been adjusted from M22 to M24.

2 METHODOLOGY

2.1 INVESTIGATION METHOD

In order to collect feedback on the different components under development from members of the different language communities, we developed a modular evaluation method that local partners could implement. The evaluation focused on three key components of the EASIER system: the application, the translation component, and the avatar. Working with the partners involved in developing these components, we designed an evaluation that included these three parts (see Deliverable 1.2: **Report on performance metrics and user study preparations** for more details). While the application was available for testing in all languages, the translation and avatar components were only available in select languages. As a result, not all language groups tested all components.

To collect end-users' input, we engaged participants in focus group discussions centred on the component in question. For each component, participants were first presented with an early version or prototype, then a discussion was led by a facilitator. This method was chosen as it allowed us to collect highly detailed, qualitative feedback that could be used to improve the components in development. Furthermore, by collecting in-depth opinions from participants across all the sign language communities, we were able to identify important overarching themes that were shared among groups and participants.

The evaluations took place over the period of September-November 2022, and included in total fourteen groups from seven sign language communities. This first part describes the user recruiting process and the procedure of these evaluations.

2.1.1 Participants

Deaf and hearing participants were recruited from the following sign language communities: British Sign Language (BSL), German Sign Language (DGS), Swiss German Sign Language (DSGS), Dutch Sign Language (NGT), Greek Sign Language (GSL), Italian Sign Language (LIS) and French Sign Language (LSF). For each of the 7 communities, there were two separate evaluation groups, one with deaf and one with hearing signers, resulting in a total of 14 groups.

2.1.1.1 Facilitators

To set up these focus groups, local partners identified facilitators for the evaluations. For the deaf group, a deaf facilitator was chosen and for the hearing group, a hearing facilitator was chosen. In two cases, for GSL and NGT, given time pressure and lack of available alternatives, hearing members of the EASIER project acted as facilitators. In both cases, these individuals were recognised as long-standing members of the sign language community (sign language researcher/CODA and sign language researcher/interpreter). Facilitators were mostly internal employees of the partner organisations, unless none was available in which case a suitable candidate was recruited from the sign language community.

2.1.1.2 Participant recruitment and profile

For each group, between 4 and 6 participants were recruited who use the target sign language. For some languages, specifically BSL and DGS, there was attention to recruiting

participants with linguistic training who would be able to read sign language glosses when evaluating the translation component. For other languages, recruitment was more flexible. Given time pressure, one group was unable to recruit 4 participants and was conducted with 2 instead.

Facilitators along with the local team were responsible for recruiting evaluation participants. For all groups, recruitment was carried out through personal and professional networks. Facilitators contacted local deaf organisations, special interest mailing lists (e.g. deaf students), interpreter associations, and reached out individually to prospective participants. Participants were compensated for their participation in the form of cash or vouchers, in line with each local partner organisations' guidelines.

The recruitment resulted in a total of 62 participants, 30 deaf and 32 hearing. Nineteen identified as men, 40 identified as women, and 3 identified as non-binary. The deaf groups were quite heterogeneous concerning profession, and included teachers, university instructors, actors, tech professionals, and students. They included participants who identified as deaf, but also as hard of hearing and included one individual who was colourblind. They also covered an extremely large age range, from late teens to late 60s, and were made up of slightly more men (16) than women (11) or non-binary people (3). The hearing groups were made up of mainly sign language interpreters and sign language researchers and ranged between early 20s to late 50s. This group was largely skewed towards women (29), with 3 men and no non-binary participants.¹

Table 1 below summarises the information on the evaluations and the conditions under which they were conducted. It recalls for each focus group which components were tested, the number of participants, and the evaluation setting.

TABLE 1: FOCUS GROUPS SUMMARY TABLE

Language	Components tested	Group	N° of participants	Setting
BSL	app, translation	deaf	4	face-to-face
		hearing	5	online
DGS	app, translation, avatar	deaf/hh	6	face-to-face
		hearing	5	face-to-face
DSGS	app, avatar	deaf	4	face-to-face
		hearing	4	online
LIS	app	deaf	6	online
		hearing	5	online
NGT	app	deaf	2	face-to-face
		hearing	4	online

¹ This is likely to reflect the demographics of interpreting, where there is a skew towards female interpreters; see for example Napier et al (2021)

GSL	app, avatar	deaf	4	face-to-face
		hearing	5	hybrid
LSF	app, avatar	deaf	4	face-to-face
		hearing	4	face-to-face

2.1.2 Procedure

Pilots were conducted for each group to allow for a dry run of practical setups and testing procedures. Feedback from these pilots was shared among the local partners to learn from each others' experiences. These pilot studies led partners to make small adjustments to ensure the smooth running of testing sessions: for example, having participants test the web app on smartphones instead of computers.

Given the scale of the evaluation and the number of partner institutions, local partners determined their technical set-up. While most elected to conduct in-person evaluations, some groups decided on online evaluations to make recruitment and participation easier. One group conducted a hybrid model where 4 participants attended in person and one via videoconferencing.

For those groups that were conducted online, participants used their own devices (either mobile phones or computers) to navigate the online components (app, avatar), and facilitators shared screens for the offline components (translation). For those evaluations conducted in person, in some cases, participants brought in their own devices and in other cases, they used devices provided by the institutions or a combination of both. In several in-person groups, facilitators also used projectors or large computer screens to provide visuals during the discussion

In most groups, the facilitator and participants were the only ones present in the room during evaluation. However, for some groups, the facilitator for the other group was also present to observe or answer questions at the end. Evaluation sessions with deaf groups were conducted in sign language, and sessions with hearing groups were conducted in the local spoken language (except for the NGT hearing group which was conducted in English). Therefore, if the deaf facilitator was present in the hearing group, a sign language interpreter was also present. The additional facilitator, and in some cases, technical staff members also assisted with video recordings to adjust camera angles, replace batteries and memory cards.

The duration of testing each component varied across groups. For the app, participants took between 20-60 minutes to explore the app, and discussions ranged from 30m to 1h20m. The translation component was relatively similar across groups, taking between 45m to 1h. The avatar questionnaire took between 20-30 minutes, and the discussion ranged from 30m to 1h 40m. Before each evaluation began, informed consent was obtained from all participants, using local partners' consent procedures. For one participant aged 16, parental consent was also obtained.

For most groups, the evaluation sessions were recorded using either video or audio recording devices. Several groups used wide-angle cameras such as GoPro's to record the entire scene. These recordings were then used by facilitators to later compile a report detailing the content of the focus group discussion. Recordings were kept by the local

institution and not shared. In two cases, given time pressure, facilitators opted not to record the session.

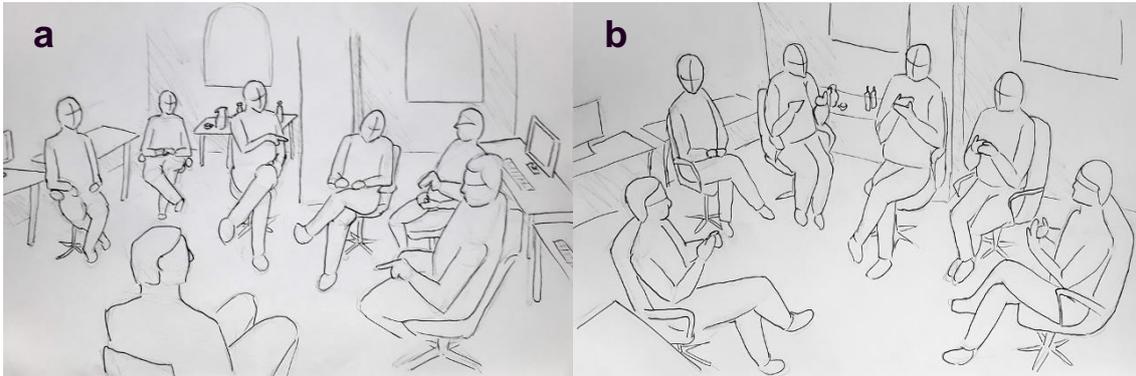


FIGURE 1: DGS SETUP FOR DEAF (A) AND HEARING (B) PARTICIPANTS. PARTICIPANTS USED DESKTOP COMPUTERS IN THE TESTING ROOM AND SAT IN A SEMICIRCLE FOR DISCUSSION WITH A LARGE SCREEN BEHIND THE FACILITATOR FOR SHOWING MATERIALS.

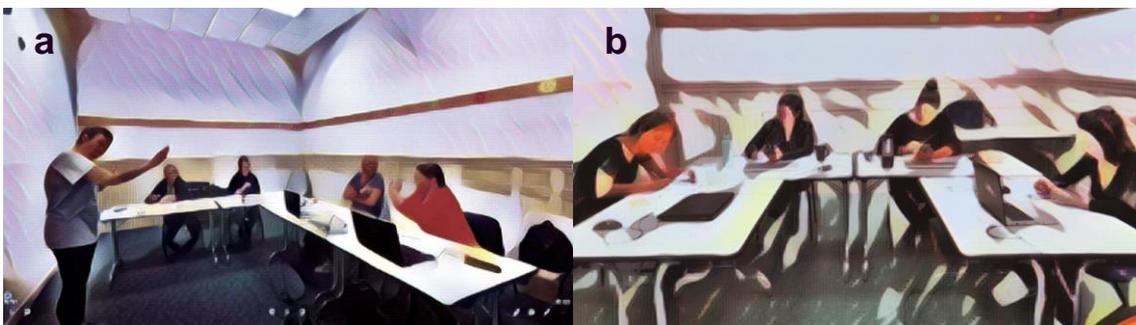


FIGURE 2: LSF SETUP FOR DEAF (A) AND HEARING (B) PARTICIPANTS. PARTICIPANTS BROUGHT USED LAPTOPS PROVIDED BY INTERPRETIS AND MOBILE PHONES AND SAT IN AN L (DEAF, LEFT) OR A U (HEARING, RIGHT) FORMATION, WITH A PROJECTOR PROJECTING ONTO THE WALL BEHIND THE FACILITATOR.

2.1.3 Data analysis

In order to compile the deliverable, local partners prepared reports on the focus group discussions which were then given to EUD. This was done for several reasons. First, it allowed us to preserve the anonymity of the evaluation participants, as they were known only to the local partners who recruited and carried out the evaluation. Aside from basic demographic information, no additional information about participants was shared among partners. Second, it allowed local partners to use their language expertise to provide English summaries of discussions in local signed and spoken languages. Given that the initial focus groups were held by EUD in international sign and Interpretis in LSF (see Deliverable 1.1 **Report on Analysis of User Needs and GUI Design**), having groups held in local languages allowed for broader recruitment and allowed participants to express themselves in their preferred natural languages. Summarising the discussions also allowed us to avoid time-consuming transcription and translation.

Facilitators followed a reporting template, and all reports can be found in the appendix. One report was delivered in International Sign, and translated by EUD to written English. The English version is in the appendix. EUD gathered all reports and compiled them into the final deliverable, and produced an analysis based on emergent themes.

3 APP

In this interim evaluation, we presented participants with an early version of the application known as a 'click dummy'. This version had all elements present, however, they were not integrated into the translation systems and therefore were not able to produce a functional translation. Instead, if sign language output was selected, the app showed a video of the avatar signing the sentence "Thank you for using our service" for the languages where this animation was available (LSF, DGS, DSGS and GSL), and for all other languages, it simply returned the phrase "Avatar Output" in the written language corresponding to the sign language selected (e.g. if BSL selected, English). (See Deliverable 8.1 **UI/UX design** for more details on the features of the app). Appendix A contains screenshots from the app.

The app was evaluated by all groups across all seven sign languages. To facilitate this, the app was localised into all project written languages, so participants were able to select their preferred language to interact with the app.

While both a web app and a mobile app are under development, this evaluation presented participants with an early version of the web app. Some groups elected to interface with the web app via a mobile phone browser to give more of a look and feel of a true mobile application, while others used a computer browser. Participants were first instructed to sign into the app, using a standard username and password combination (user@example.com; user). One group performed their evaluation after an app update, and were able to register to the app with their own email address. After signing in, participants were encouraged to freely explore the app's features before returning to the group for a discussion.

In the following chapter, we will first present an overview of participant feedback regarding the app experience (section 3.1), then present detailed feedback on different issues of the app interface (section 3.2), and end with some feedback on the evaluation protocols regarding app evaluations which we can then use for future evaluation phases (section 3.3).

3.1 GENERAL FEEDBACK AND RECOMMENDATIONS

3.1.1 App navigation

The navigation in the app turned out to be one of the major points of dissatisfaction during the evaluations. An overwhelming majority of evaluators, both deaf and hearing, pointed out that they had trouble navigating through the app, that the navigation was not intuitive and took too many steps/screens. Particularly the need to press the “Back” arrow was strongly rejected, as this is not an intuitive linear progression.



FIGURE 3: LOCATION OF THE BACK ARROW IN THE UPPER BAR

Users were very critical of the progression throughout the interface, citing it as not user-friendly or even as “*painful*”. They complained about the fact that it was not linear and required constant steps backwards, particularly when picking the input and output modalities and languages. Currently, there is no automatic progression when selecting options, and it is unclear if settings were saved. Some said they expected some kind of visual feedback after selecting a setting. The current buttons feel “dead” and there is no feedback to tell them that the option has been selected, which in some cases caused multiple attempts before realising that the option was actually already selected and that they simply needed to return to the previous screen. Users expect to be automatically driven to the next step, only resorting to using navigation arrows if the user specifically wants to go a step back or forwards in a linear progression and to have a “confirm”, “save” or “OK” button to clarify that the settings were saved and the machine translation can now be used.

Participants suggested a “Home” button, which brings one back to the landing page.

3.1.2 Need for guidance

Users consistently said that the application did not feel very intuitive to them; they had difficulty in both understanding what it could be used for and how to use it. They observed that the app provides little information about its use and usefulness. This feedback shows that the affordance of the application, i.e. the ability of an object or system to evoke its use and function, is insufficient.

Due to the navigation being labelled as “unintuitive”, participants also remarked they needed more guidance on how the app is supposed to work. Some specifically remarked that they needed a “first-time user” manual of some sorts, where an overview or tutorial is given. Some step-by-step guidance would be appreciated, as users were often left with the question, “What am I supposed to do now?” Some people suggested a FAQ screen, and others suggested a video manual on how this app works. As the app is now, you have to make guesses yourself and try to figure out how it works.

Participants also wondered who to contact in the case of questions or issues, such as where they could go if they forgot their password.

3.1.3 Translation screen and language selection

Within the current interface, a majority of the evaluators faced some kind of difficulty while selecting the input/output modalities and languages for the machine translation. One participant stated that *“the application requires multiple back-and-forth clicks which is very time-consuming and confusing”*. They also found it confusing to only have spoken languages as input languages and sign languages as output languages. Some said it seems like they have to have only one language for all of the three options, since it shows the same language selected, even when the input modality is changed.

What bothered deaf users was that the language selections didn't show the names of the different sign languages. Upon selecting the camera as an input modality, in the next screen the spoken languages are named: e.g. “Italian” instead of “LIS”, or “Dutch” instead of “NGT”. Evaluators felt their sign languages were not really represented in the app.

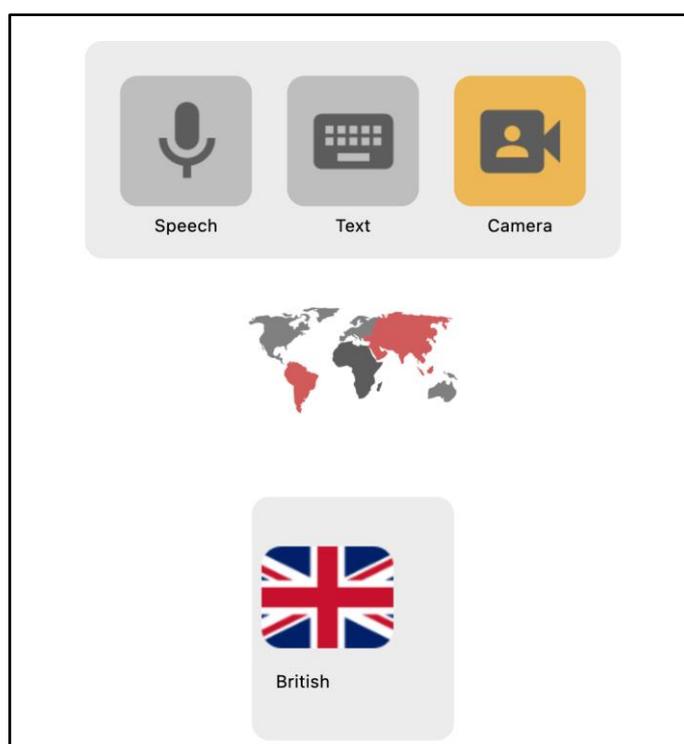


FIGURE 4: “BRITISH” IS SHOWN AS SELECTED LANGUAGE FOR AN INPUT THROUGH THE CAMERA

After finishing the first translation, there should be a way to go back to the input screen and add a new input, instead of being thrown back to the start of the whole process and having to start all over again with all the selections. They also criticised not being able to go back to change the input itself, e.g. in the case of spelling mistakes, as well as input and output language and modality. Several participants said that they felt the selection of default language input/output was redundant since every attempt to perform a translation seemed to ask them to confirm the language selections again. They felt that your last used preference should be remembered and defaulted to.



FIGURE 5: “BACK TO START MENU” BUTTON AFTER DOING A TRANSLATION

Regarding the translation interface, evaluators often referred to well-known machine translation tools such as Google Translate or DeepL and preferred a similar design for the EASIER application, i.e. having the input and output next to each other in the same screen instead of on different screens with multiple steps in-between. There was one discussion that these elements next to each other might pose a problem on a small screen, but that one above the other was thought to be a good solution. The tool being on one screen is very straightforward to use.

It was noticed that some users would prefer a horizontal use of the mobile screen since they are strongly influenced by the boxes providing input and output in Google Translate. They would like to have a similar screen arrangement in the EASIER app as well. This was especially preferable in the case of translation from one SL to the other.

This also applies to the settings for the input/output languages and modalities, these should be on the same screen, as users do not like the dissociation of the input/output settings. The current deviation of the common standard as used in popular machine translation tools is seen as contrary to user-friendliness. Participants suggested drop-downs for language and modality selection for both input and output within the same screen, along with a button to quickly swap translation directions, to allow users to change the input or the settings there and then. This enables users to save time in navigating the app. Participants also suggested using clearer icons for the different language modalities.

One proposal was also to let the voice style setting of the output be decided at the same time as the language choice, for both spoken language and sign language output.

After the translation process, the users were not clear on the exact functionalities of the “Copy”, “Share” and “Save” buttons. Participants felt they needed more information about the differences between the buttons. They had questions like “*Where will the files be saved?*”, “*Why should I copy? Why sharing?*”, “*Can the saved data be deleted?*”



FIGURE 6: “DELETE”, “COPY”, “SAVE” AND “SHARE” BUTTONS AFTER DOING A TRANSLATION

3.1.4 The use of flags

The flag icons in the translation screen do not specifically mention the selected language. For example, there are three written languages in Switzerland, but it is not possible to deduce from the Swiss flag icon which one of the three was selected.



FIGURE 7: SWISS FLAG ICON DISPLAYED AS INPUT LANGUAGE FOR A TEXT TRANSLATION

Some people were surprised that the flags could not be clicked on, but that they had to click on the small buttons with the names of the languages underneath. They would have

preferred to be able to click on the flags, however, this does not really make sense when there are multiple language options in one country. Another comment made by some participants was that people might possibly not recognize all flags or thought some choices were weird (e.g. the choice to put Luxembourg as a category), so it is recommended to add country names as well, or that a list of language names is added instead of flags, which would be more convenient to users and takes up less interface space.

3.1.5 SL Video recording and minimum quality requirements

The evaluators had different comments, questions and thoughts about the video recording.

- ➡ Is there a time limit for video recordings?
- ➡ Is it possible to record entire conversations or just bits of it?
- ➡ Is it possible to record someone else signing? (using the back camera)

One participant asked if there were any requirements for the video recording. How does the user know when the video is of good quality for translation? E.g. wearing fingerless gloves because it is cold, having half the face hidden behind a shawl, wearing big earrings, etc. Will the app understand the signing, or are there certain requirements that need to be met, so that the app will understand the user well? There is currently no indication.

Other participants wondered if one-handed signing would work with the machine translation and how much flexibility they have for one-handed signed input (for example, using different objects/body parts as substitutes for the hand holding the phone). They also wondered if they could add to the already existing video recording by overwriting only a part of it.

For video input, the participants also requested visual feedback to indicate that the recording started/is in progress and maybe also how long the recording can be. They imagined something like a bar showing that one minute can be recorded and time is running. Another participant had the idea for an input check, that the video recording screen could have a coloured frame with green meaning that the angle, light and distance from the camera are okay, red meaning that the input is not good, that the angle, light or distance should be changed.

Participants also mentioned that the video screen is too narrow to record the whole signing space. They missed the function of a double click to change the video screen into full screen, as the button to change into full screen is very small. Participants did not like that this option did not widen the camera angle. Some participants would like to turn their phones horizontally to record videos. The screen format of the FaceTime app was named as a good example to follow for the recording screen.

Participants also found it irritating that the input video was not mirroring the image of the signer, which is a practice used widely. Placing oneself correctly in the frame of the camera was a bit difficult as it is counter intuitive when the video is not mirroring the signer, as the signer then intuitively moves in the opposite direction. They suggested that this could be an additional option in the settings; to have the video mirrored or not.

They also discussed the buttons in the recording screen. Some participants found it difficult to find the “record” button, as it is too small and not placed intuitively (inside the screen). It was suggested that the button to start the video recording should be red and maybe beneath the video screen, which is a common practice (e.g. the iOS camera interface). They would also like to be able to pause the recording when being interrupted, and to play the recorded

video to see if it is as intended. Participants suggested using the commonly known video buttons “Record”, “Pause” and “Play”.

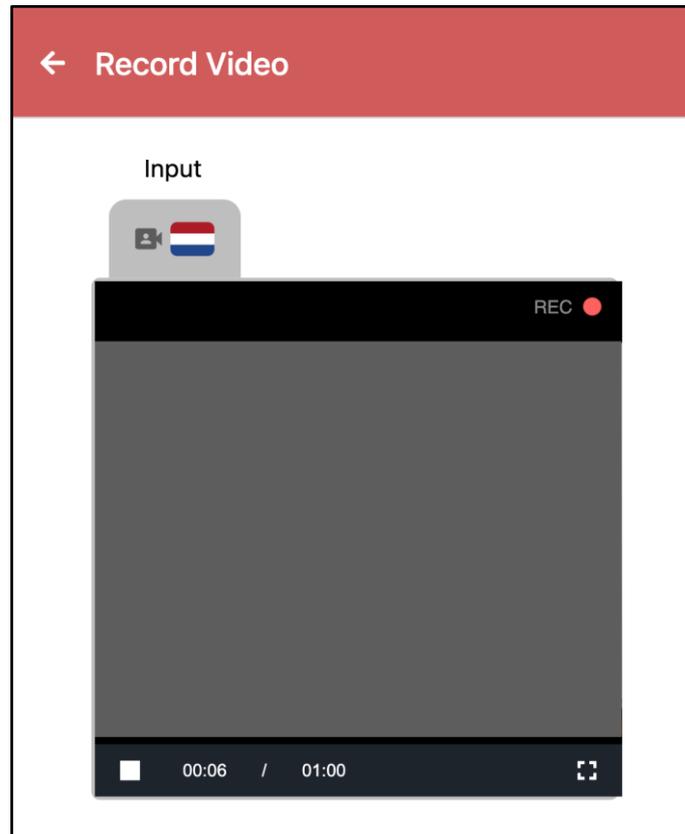


FIGURE 8: RECORD VIDEO SCREEN

3.1.6 Audio recording

Participants wondered if speaking into the camera was possible for the machine translation, or if only signing works for the video input method. In the first case, both spoken and signed languages should then be included in the language selection settings. Another question was what happens if while speaking single signs are used.

For consistency, one user mentioned that the same icon should be used all over the app for vocal communication and voice recording, i.e. the microphone.

The question arose if there will be some visualisation of the voice being recorded. The participants would like to see some kind of visual feedback that the recording is happening. At least one participant could not see if the app was recording or not, but her phone told her that the microphone was in use. This also led to the question if the microphone and camera are active permanently while using the app. Participants raised the concern that this would mean high power consumption.

3.1.7 Text input

Participants discussed the quality of the written input. They wondered if the app will be able to translate the input in cases where there are spelling mistakes or wrong use of capitalization. They suggested having a spell-check and underlining erroneously spelled words so that users can spot mistakes faster. Another suggestion was to have suggestions

for synonyms and a list of possible words if the word is spelled wrong, e.g. when the written word is unknown to the machine translation.

Participants also wondered what happens when languages are mixed in the input, e.g. “Denglisch” or loan words in the input, or when gender-neutral language is added in (with e.g. asterisks or colons, for example, *Freund:in* or *Freund*in* in German).

One participant had the idea for a fourth input method: via the picture of a text. She compared this function to Google Translate and mentioned that this would be extremely helpful to deaf people. The group strongly agreed with this idea. The menu in a restaurant was named as a use case, as it is an extremely quick way to translate written text.

3.1.8 App settings and clarity of icons

Participants made positive remarks about the options for the avatar output, (gender, clothing and background colours). Still, they would have liked to see a visual example in action on the selection of the different options. Some did not understand the difference between “Background” and “Contrast” and felt that if the avatar example had reflected those options upon selection, they would easily have understood the difference. Some thought the colour options for the clothing could be reduced to just a white or a black shirt for optimal contrast with skin colours.

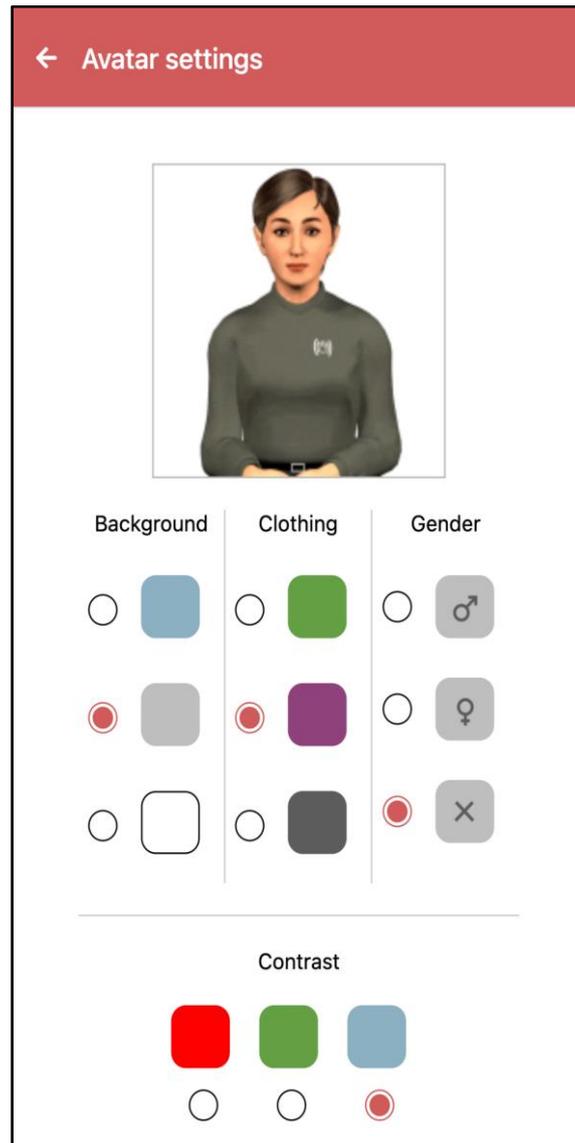


FIGURE 9: AVATAR SETTINGS SCREEN

Participants did not like that they had to tick “Use favorite settings” to have their settings saved. They would like that to happen automatically. For some users, it was not clear what this option referred to. They did not understand whether it referred to the communication modalities and language choices, or also the avatar and/or contrast settings. One user mentioned having been confused by the “double” settings functionality. He did not understand right away that he first had to go into the settings in the drop-down menu to be able to select the “use favorite settings” option afterwards.

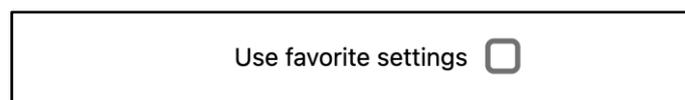


FIGURE 10: “USE FAVORITE SETTINGS” OPTION IN THE START SCREEN

The symbols in the icons were not clear to everyone. Some people thought the voice style icons for formal (suit) and informal (t-shirt) were actually choices between male and female. Some thought it was for the avatar’s look. Others understood the meaning but disagreed with

the suit since this is not gender-neutral, but leaned towards male representation. Others were confused about what “voice style” meant. One user asked if this setting also applied to the sign language of the avatar. “Adapt intonation” was also not understood. Participants found these settings particularly not self-explanatory.

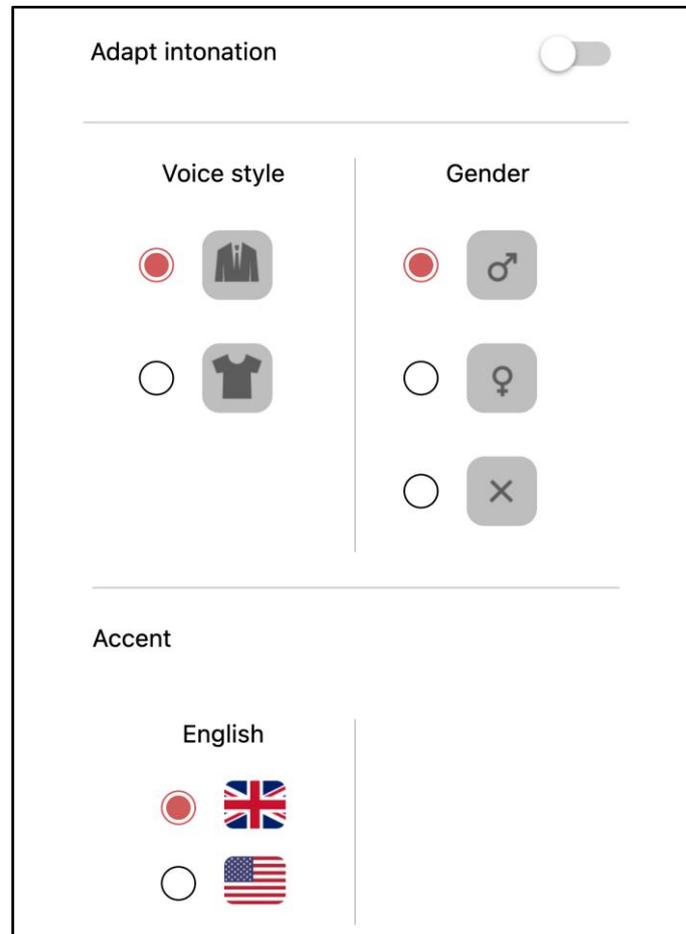


FIGURE 11: VOICE SETTINGS SCREEN

The camera icon in the translation options was also a bit confusing for some of the evaluators. Some assumed for example that speaking in the video recording could be used as audio input for the translation. The suggestion to use a sign language pictogram was made. Participants wanted clearer icons for the different language modalities.

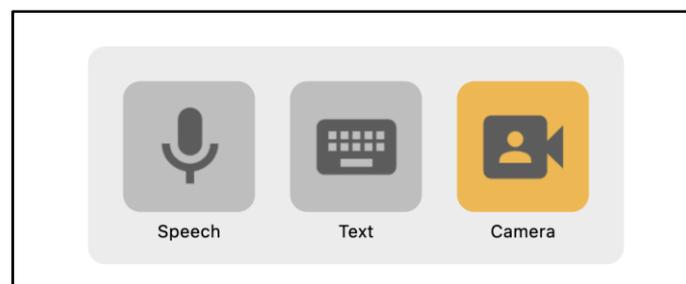


FIGURE 12: INPUT MODALITY SELECTION

Evaluators expressed a preference for having easy access to settings, possibly through a visible settings button in the menu. Some participants did not find the interface language settings until the option was pointed out.

“Default input” and “Default output” settings were not understood by all, some thought they were for peripherals such as microphones, speakers, etc. It was not clear that those settings were referring to the language selection.

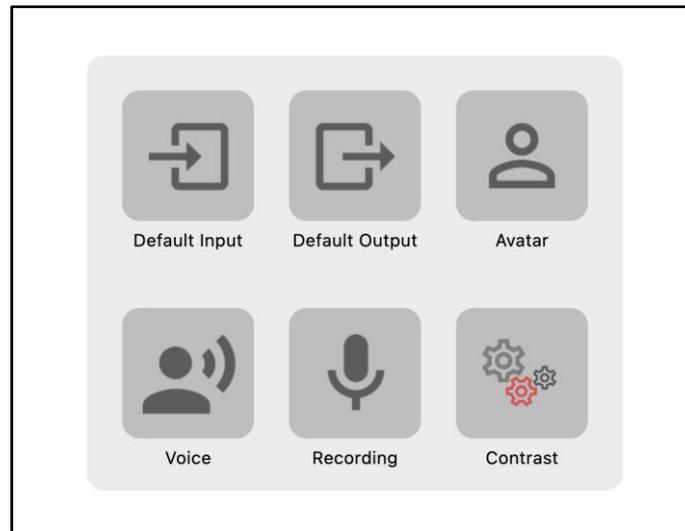


FIGURE 13: OPTIONS IN THE SETTINGS SCREEN

3.1.9 The voice of machine translation

Participants also expressed concerns about how the voice is linked to the avatar. While choosing the avatar, they should also be able to choose a voice, or while setting up default settings, such as language, the app should be able to use that information to set up the voice to meet the user’s age and characteristics.

Issues with the accents were mentioned. The question about the American vs. British accent option was brought up. Why is there an American accent option, as this is an European project? What about the different German accents in Europe, or even regional accents in the United Kingdom, such as London, Yorkshire, etc.?

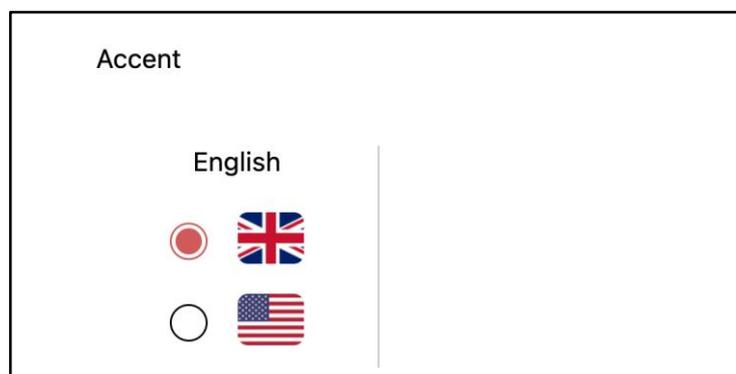


FIGURE 14: ACCENT OPTIONS FOR SPOKEN ENGLISH

How would the app adapt for intonation, to appear interested, aggressive, gentle, polite or angry? At the moment there is a toggle in the settings - but what would that sound like? Deaf people have no idea what the different settings do. They would like to have examples for gendered voice settings.

3.1.10 Dark mode and accessibility

People liked the option to change the app in a dark mode, however, they noted several technical and visual issues with the dark mode, and some had recommendations to further improve the dark mode and/or related accessibility settings:

1. There should be an accessibility button immediately on the front page. Easily findable accessibility settings are usually on the top left or right corner of any application or website. The current path to the accessibility settings could be improved, as participants had some difficulties with finding the setting, especially those who actually need the accessibility settings the most.
2. In conjunction with the possibility of increasing font size, also offer the option for font thickening. Some stated that they would like the yellow font colour in the dark mode to be white, or that it should be optional - there was the suggestion to let the users choose. The comparison with Netflix subtitles was made, where you can choose the text colour, size, boldness, for an optimal viewing experience for the user. *“Don’t force one choice to the users”* as was said in the discussions.
3. One participant remarked on the colour selections and wondered if they were colourblind-safe.

NOTE: Specific issues with the dark mode are listed in the next chapter, Issue-specific feedback.

In a discussion, the topic of the target group was raised. Participants discussed that there are different groups ranging from children to seniors and that the app should meet a certain average to be appropriate for everyone. Problems estimated for seniors were the size of the screen, the intensity of mouthings on the avatar and the use of fingerspelling. It was also stated that seniors will probably not be the main target group. It seems that younger generations understand the app better when compared to older participants. They appear more flexible and know how to use the app, so the broad skill sets of different generations should be considered.

Some participants also wondered if the avatar needed a logo on her chest. Maybe it could be impeding the visibility for some. A good option would be to have the logo in an upper corner.

3.1.11 Representation of diversity

Participants to the evaluations had issues with the representation of diversity in the app. There should be more diversity than just a white male/female/non-binary avatar, other options named were: size, age, “race”². If deaf people who are not white see a white interpreter, they don’t feel represented and they feel excluded. That there is no possibility to choose different skin colours can be perceived as an ethnic discrimination. A quote from one of the participants: *“Don’t forget that there are BIPOC people in Europe as well!”*³

² “Race” as said by participants. This includes skin colour, hair texture and other physical features of ethnic minorities.

³ The term BIPOC stands for Black, Indigenous and People of Colour. While it originated in the United States, it is sometimes used in other contexts to describe people of ethnic minorities in majority white countries. In this case it was used by Italian participants.

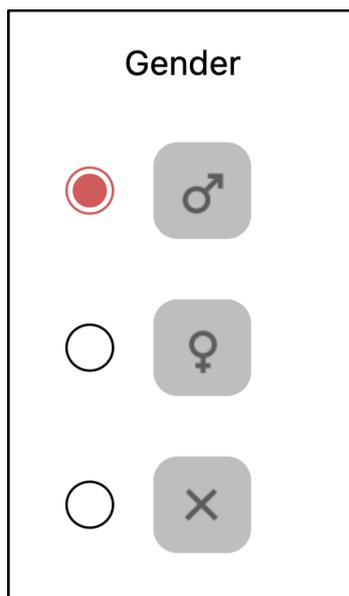


FIGURE 15: GENDER OPTIONS IN THE SETTINGS

While it is viewed positively that the gender options are not limited to only male and female, some participants didn't like the gender icons used for male, female and non-binary in the settings, claiming that grassroots deaf people might not understand them. Furthermore, the non-binary icon, which is pictured as "X", was especially criticised because it's considered to be depreciative and stigmatising, or could even be seen as a discrimination because the "X" is perceived as a negative or bad thing. For example, intersex people did not fit in with any of the options. They preferred "other" or a different image. Another participant emphasised that the "X" is read by her as clearly representing agender people and that the official non-binary symbol is more inclusively representing everything that is not male or female. Some said they preferred just words over icons.

Gender-neutral language was also brought up, it would be nice to have the option to have gender-neutral language (or different styles for gender-neutral language) for the machine-translation output. One participant criticised that the info text is not written in gender-neutral language. This info text also does not mention "deafblind" alongside "deaf" and "hearing".

3.1.12 Interaction flow

Deaf persons mentioned that the way the app is currently constructed constrains the shape of the interactions between the interlocutors and does not allow a smooth dialogue. On the contrary, it produces an interaction that feels like a consecutive translation instead of a simultaneous one like human interpreters do.

The progression through the interface of the app implies that the user first sets the input modality and the language. This first action leads directly to the recording of the message, once that has been recorded, the user has to set the output modality and the language to be able to generate the translation. The translation is then shown to their interlocutor who goes through these steps again to respond. This process is visualised as follows:

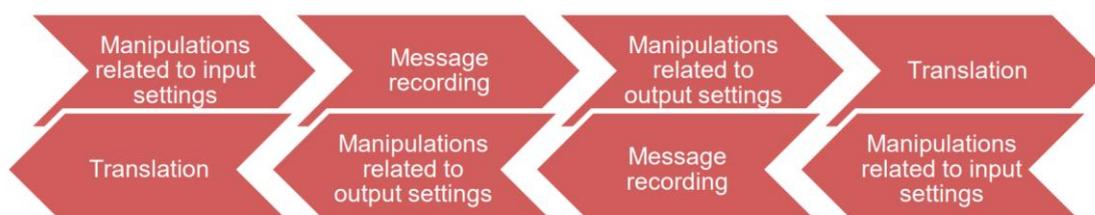


FIGURE 16: VISUALISED PROCESS OF AN INTERACTION FLOW BETWEEN INTERLOCUTORS WHILE USING THE CURRENT EASIER APP

The above image clearly shows that the natural sequential organisation of a conversation is disrupted by the manipulations regarding the input/output settings. Deaf persons mentioned that these steps increase the sequentiality of the interaction, interfere with the fluidity of the conversation and remind them of conversations written on pieces of paper.

This leads to a feeling that the use of the app can be a hassle. When a deaf user arrives at a location and has to go through all these steps, it takes too long to actually do the conversation, they may want to revert to “default” communication such as pen and paper, gesturing or other methods such as the Big Word app.

The recommendation is to simplify the process to get communication going, i.e. minimise the time needed to start the input of a message and the translation thereof. Participants in the evaluations indicated with an overwhelming majority that the current navigation takes too many steps and clicks/taps.

They envisaged the use of the app out on the streets, having to have a back-and-forth exchange with someone, most likely passing the phone back and forth. They felt that the app should be striving to make that exchange as smooth as possible, by minimising the number of steps and clicks/taps.

3.1.13 Archive, data security and privacy

Users had questions about the protection and the archiving of their data. How long is personal information stored? Why is the ‘archive’ there? Who maintains this data? Can the saved data be deleted?

Participants also wondered where their data was being stored as it could be sensitive content. It was discussed that if data is saved in a cloud server this could be a problem with data protection, but that locally saved data uses up a lot of space on the device. The participants would like this information to be transparent.

Files in the Archive should be named by keywords from the content that was translated, instead of numbers.

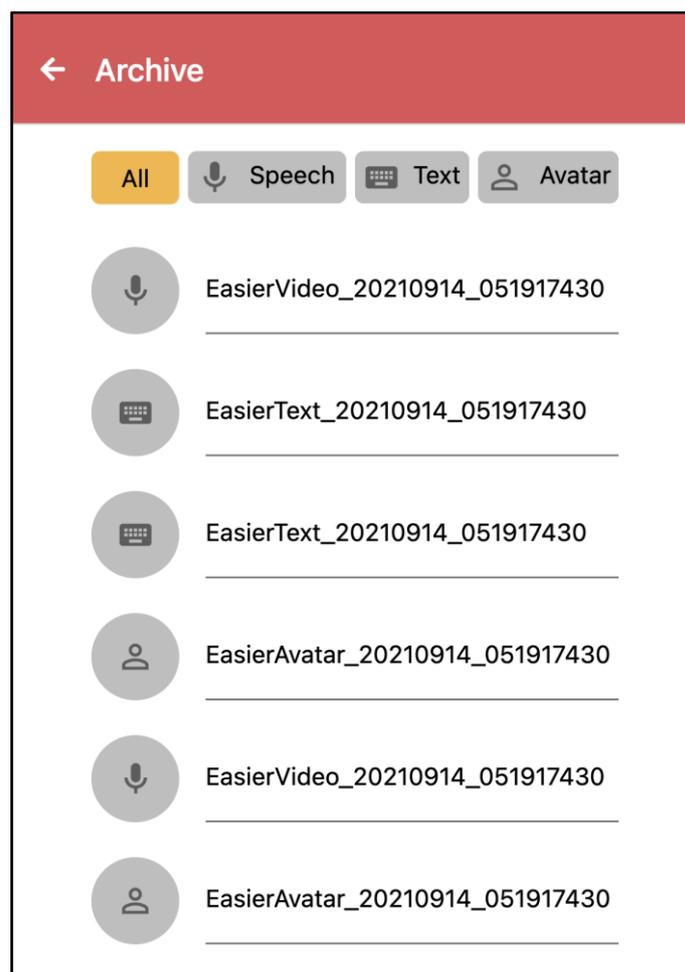


FIGURE 17: ARCHIVE SCREEN

One participant had the idea to have the different inputs visible as a history comparable to a chat inside the input window.

3.1.14 Profiles, user IDs and logging in

Some users referred to other applications they are used to and pointed out that these applications usually link their communication preferences to the “Profile” settings. They would also like to be able to save multiple profiles to quickly adapt to different situations, as they find the whole settings process time-consuming and complicated.

Participants mentioned that the feature customisation of the avatar should be done during the creation of your user profile, and also have the default communication preferences set up during this period.

Some felt that the profile was very basic, as there are only three settings: username, email and password. There should be preferred pronouns, race, age, etcetera so that the app can represent the user. They stressed that the registration should be easily accessible though.

Participants criticised that they had to open the website again after accidentally closing it and each time they had to enter the login credentials again, which they did not appreciate. They would prefer to have this information saved for further uses.

3.1.15 Library of commonly used sentences

Some participants had the idea to have a “library” of typically used sentences readily available in the application, such as “*Where is the nearest train station?*”. This would decrease the amount of time the users needed to spend on the input. This library might be a standard library with sentences which are already pre-installed in the app, and/or “favourite” sentences that are set by the user. These sentences could be organised into categories. One of those prepared texts should be an explanation about the app itself for interlocutors that are not familiar with the app.

3.1.16 Choices in image design

Comments about certain choices in the design of the images in the app surfaced. Some people were confused about certain design choices or looked for a meaning behind the images.

Take for example the world map image in the language selection screen, especially this image was often cited as confusing or misleading. The areas in red or grey on the world map did not match the available language options, and there were comments that the choice for a world map was strange since EASIER is an European project, with only European languages.

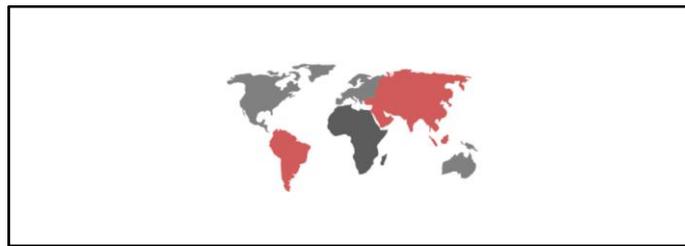


FIGURE 18: WORLD MAP IMAGE USED IN THE EASIER APP

Participants also wondered about the presence of Asian text characters in the language selection button - why are there Asian characters, as this is an EU project and those languages are not included? Does it mean that Asian languages are also an option to translate to?

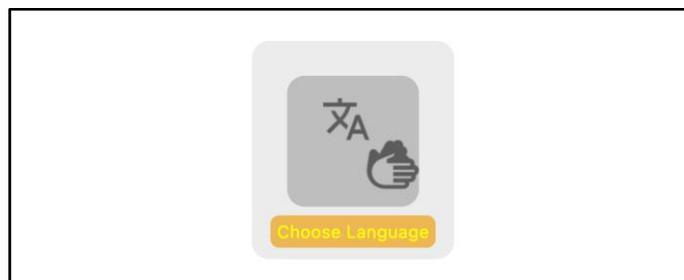


FIGURE 19: CHOOSE LANGUAGE BUTTON IN THE EASIER APP

Some liked the image with the translation from the question mark to the phone and vice versa, but others were confused by it and thought they should shake or rotate their phone. The image of the smartphone also can be misleading, as the app is also available to use on a computer.

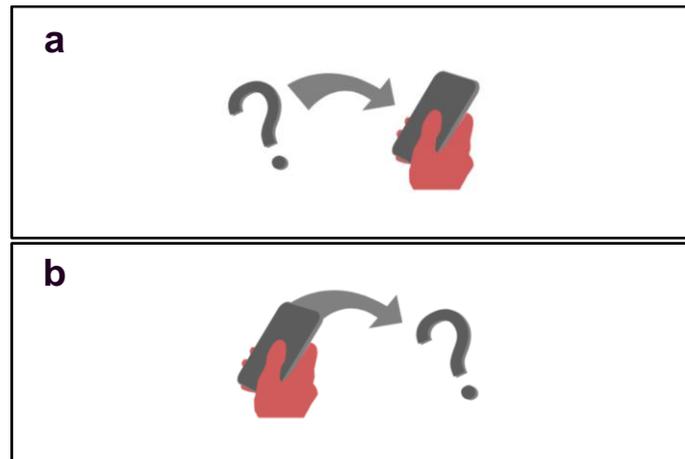


FIGURE 20: IMAGES USED IN THE EASIER APP

Some participants saw the recurrence of the colour red over and over in the images through the translation process and kept looking for a meaning behind the colour red. They felt like it meant something but could not figure out what.

In the settings, there are also two identical images, one as art/background for the settings page and one as icon for one of the settings categories (Contrast). This confused the users.

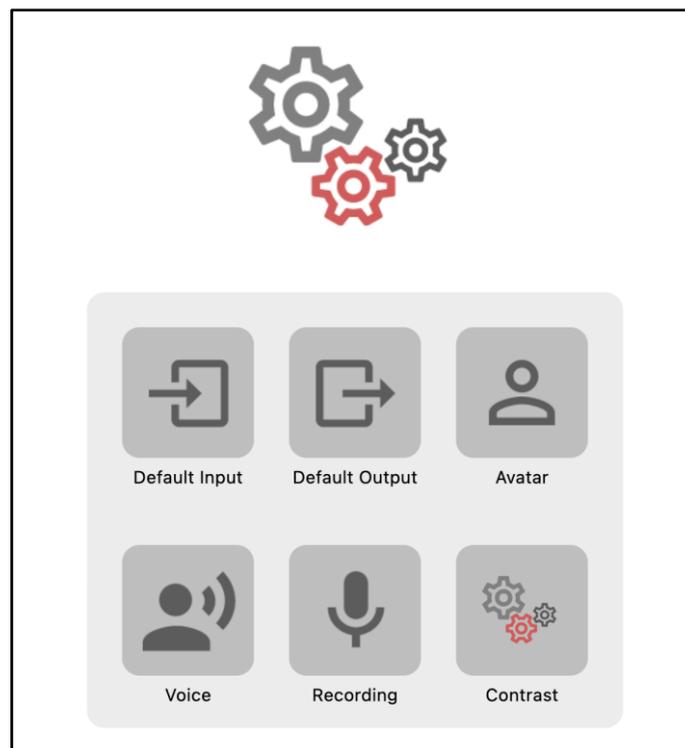


FIGURE 21: SETTINGS SCREEN

3.1.17 Offline availability

Participants wondered if there will be an offline app that one can download beforehand (e.g. when having access to Wi-Fi). One participant compared this to a dictionary app she uses which offers the service to download single languages.

3.2 ISSUE-SPECIFIC FEEDBACK

3.2.1 Localization issues

- ➞ Language options that were localized in Dutch are not in Dutch but in a Scandinavian language (ending on -sk).
- ➞ In some cases the language options stayed in English.
- ➞ The localized language options follow the alphabetical order in German.



FIGURE 22: LANGUAGE OPTIONS IN ANOTHER LANGUAGE WHEN DUTCH IS SELECTED

- ➞ Some participants found untranslated terms in English in the localized versions of the app, e.g. the “Next” and the “Start” buttons.
- ➞ Some also identified literal translations from English that did not make sense in their language.
- ➞ While not wrongly translated, the Dutch localisation has a Flemish flavour to it.
- ➞ Other suggestions were made to clarify certain app elements:
 - In French:
 - “Sexe” should be replaced by “Genre”
 - “Text clair ?” should be replaced by “Supprimer le texte ?”
 - “Accentuer” should be replaced by “Accent”
 - “Méthode d’entrée” should be replaced by “Mode de saisie”
 - In German:
 - “Aktiviere Kammera” should be replaced by “Aktiviere Kamera”
 - “Kopierne” should be replaced by “Kopieren”
 - In Italian:
 - The sentence “Toccare e scrivere” sounds weird in Italian, we don’t use “touch” so it’s better leaving just “Scrivere”
 - “Uscita di traduzione” has no meaning in Italian, it’s a calque.

- Translation issue: the translation says “is the text clear?” upon clicking “Clear” in the text input screen, but the real question is “Are you sure/do you want to continue”? The icon with the red X, is not coherent from an iconic point of view.



FIGURE 23: “CLEAR TEXT” CONFIRMATION SCREEN IN ITALIAN

- The design is not adapted to the different word lengths from the translations. Some words overlap the rectangles, other words are broken up into two lines or overlap each other.



FIGURE 24: WORDS BEING BROKEN UP IN THE GERMAN VERSION OF THE APP

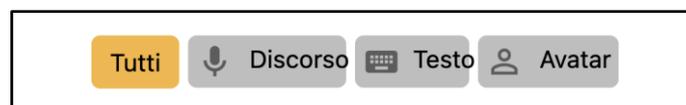


FIGURE 25: WORDS OVERLAPPING DESIGN IN THE ITALIAN VERSION OF THE APP

3.2.2 Navigation issues

- The language choice screens do not match with the chosen modalities, they are mixed up:
 - When selecting the language for the camera input, the app only proposes spoken languages. Sign languages are therefore missing.
 - When choosing the written language for the translation output, the app suggests sign languages, which is a mistake.

- On the other hand, when choosing the avatar's sign language, the app suggests spoken languages, which is also a mistake.
- ➔ Some participants were irritated that they had to click exactly the “back” arrow button, but could not click on the word “back” next to it to navigate back.
- ➔ There were issues with flag icons and language codes beneath them (e.g. the German and Luxembourgian flags show “DSGS” beneath it, this should be “DGS”).



FIGURE 26: LANGUAGE SELECTION SCREEN WHICH SHOWS DSGS INSTEAD OF DGS

- ➔ When using the “Backwards” button of the phone browser’s internal operating system the app is closed (as it is only a website and the back button of the phone puts one to the last opened website before the used one). Participants found this very frustrating and discussed that in addition to the “Backwards” button within the app they would like their phone's internal “Backwards” button to also work.

3.2.3 Design issues

- ➔ Activating the camera initially leads to a microphone icon; this was noted to be confusing and not culturally appropriate.

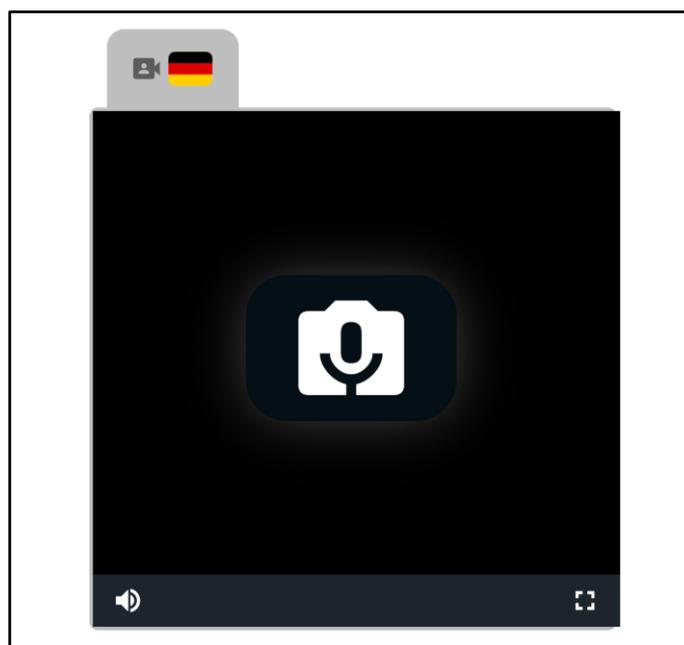


FIGURE 27: VIDEO RECORDING SCREEN DISPLAYING A MICROPHONE ICON

- ➔ The language menu is not entirely visible on mobile. It's also in a smaller font.
- ➔ One participant noticed that the avatar is signing outside the screens so that a finger is cut off.

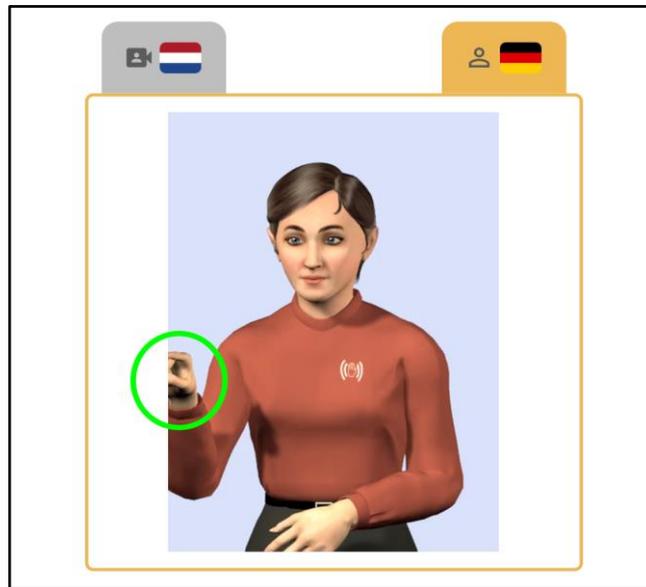


FIGURE 28: SIGN LANGUAGE OUTPUT SCREEN WITH THE AVATAR SIGNING OUTSIDE OF THE SCREEN

- ➔ Participants criticised that not all buttons are immediately visible on the screen, but that they had to scroll down to be able to use the “next” and “start” buttons. They thought that the app should adjust itself to the screen format of their phones.
- ➔ There is too much white space on the desktop version, and people wondered why the design was not optimised.
- ➔ There was a suggestion for the background colour of the buttons to change when you hover over it with the mouse, as visual support.
- ➔ There are also issues with the alignment of some design elements. For example, in the Settings screen, the icons for input and output are not aligned properly.
- ➔ The order of the icons is semantically wrong, it would be better to have “Voice, Avatar and Recording” on the same line.

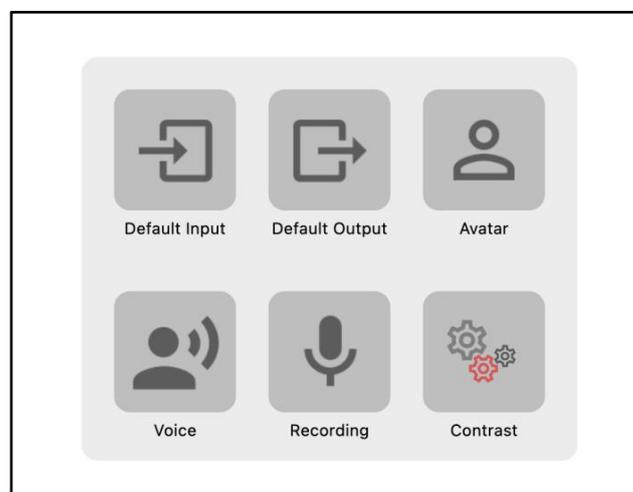


FIGURE 29: SETTINGS SCREEN

- ➡ The name of this line of icons is not semantically correct. Or you choose the “device” or you choose the “function”. So either you put “Speaking, writing, signing” or you put “Microphone, text, camera”. But this mixing is not good.

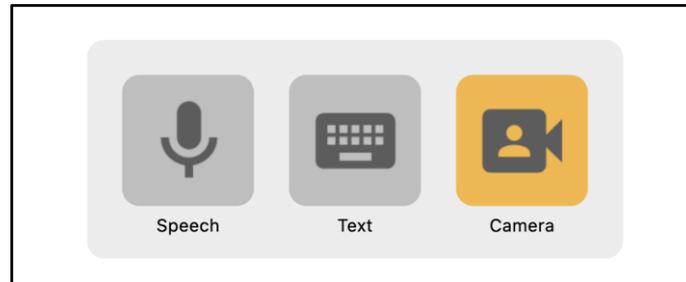


FIGURE 30: INPUT MODALITY SELECTION

- ➡ The starting page looks simple enough, but there was some confusion about having both the word Start in the upper left corner, and the Start button.
- ➡ The drop-down menu is open when the landing page is opened. This was very irritating to the participants.
- ➡ They also reported the bug that when the drop-down is closed and opened again, it shows a subcategory of the listed points.
- ➡ Some participants would like the drop-down menu to be in a different colour than the background.

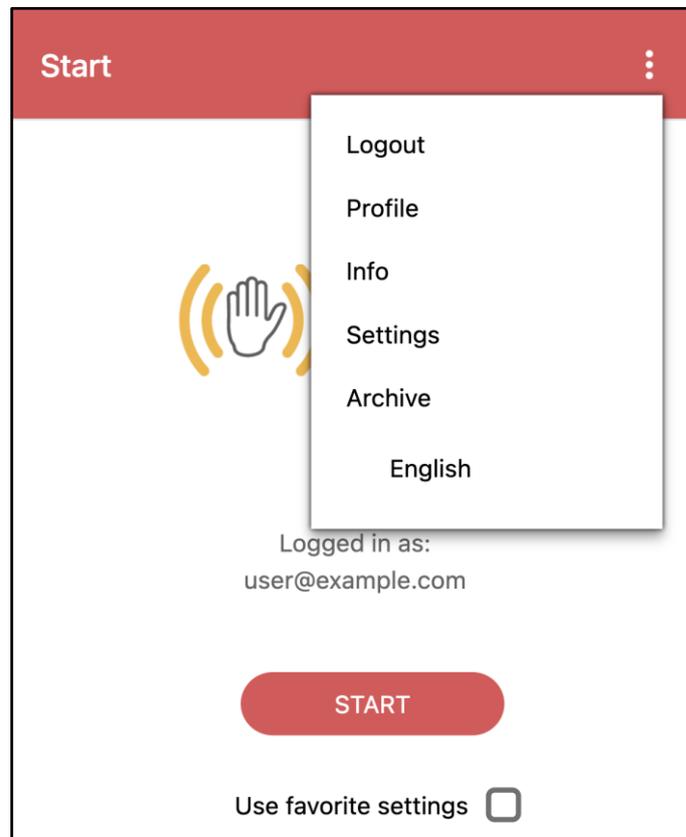


FIGURE 31: START PAGE OF THE EASIER APP WITH THE DROP-DOWN MENU ALREADY OPEN

3.2.4 Dark mode issues

- ➔ Some stated that the dark mode in the app works fine but that when the device itself is already in dark mode, the app is not shown properly, e.g., the upper bar is not distinct from the screen anymore. This means that the device itself has to be changed to light mode, and then the app itself is set to dark mode, which seems rather complicated.
- ➔ Many areas/buttons are not distinct anymore in the dark mode.
- ➔ The images are not adapted for the dark mode, so the greys and blacks are (almost) invisible when the dark mode is enabled, e.g., the black stripe of the German flag seems to disappear because there is no contrast with the black background, or parts of the logo are missing in the dark mode.
- ➔ In the dark mode, input and output language choices are not visible because of the bright yellow thin font appearing on a grey background. This contrast should be improved.

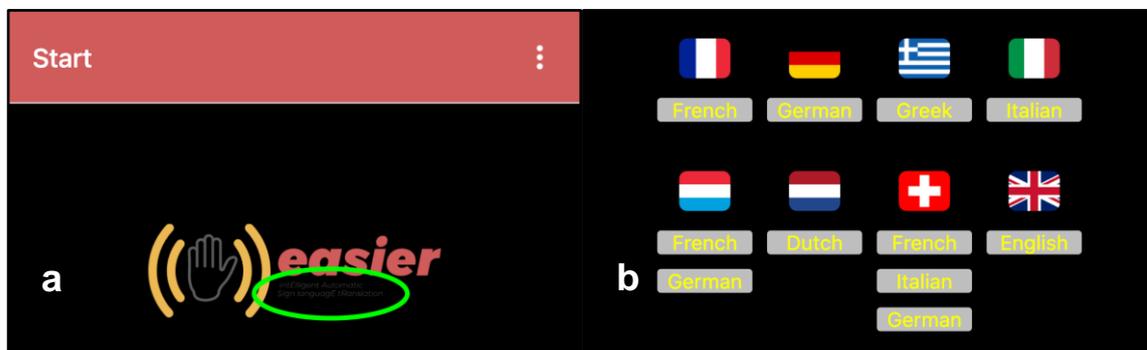


FIGURE 32: GREY TEXT IN THE EASIER LOGO IS BARELY VISIBLE IN DARK MODE (A), BLACK STRIPE IN GERMAN FLAG SEEMS TO DISAPPEAR AND THE YELLOW TEXT ON A GREY BACKGROUND IS HARD TO READ (B)

- ➔ When the dark mode is enabled, the chosen language is no longer visible in the drop-down menu, or the drop-down menu appears in white.

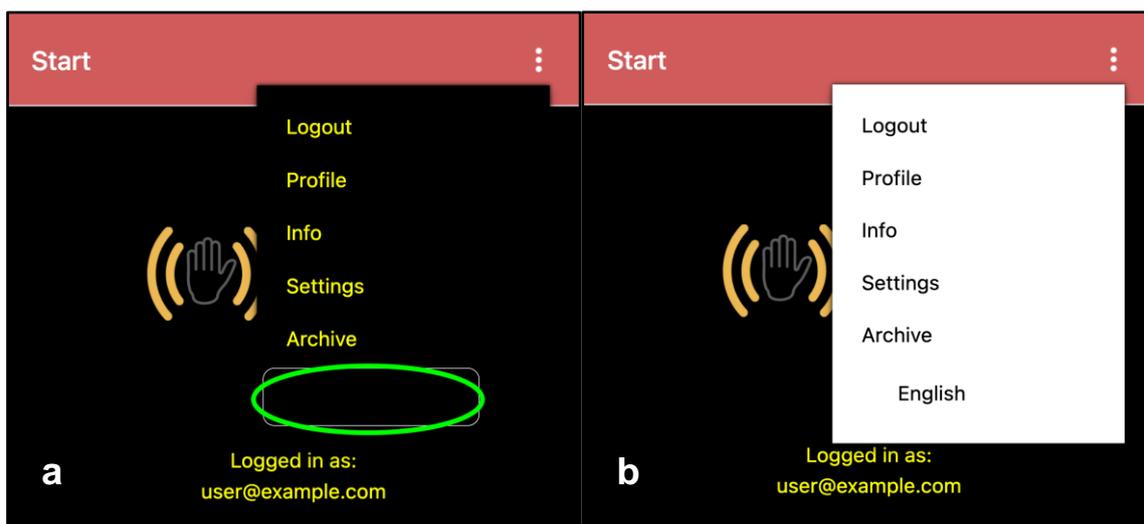


FIGURE 33: CHOSEN LANGUAGE IS NOT VISIBLE IN THE DROP-DOWN MENU IN DARK MODE (A), DROP-DOWN MENU APPEARS IN WHITE (B)

- ➔ The language menu in the drop-down appears in white, which is bothering users using dark mode.

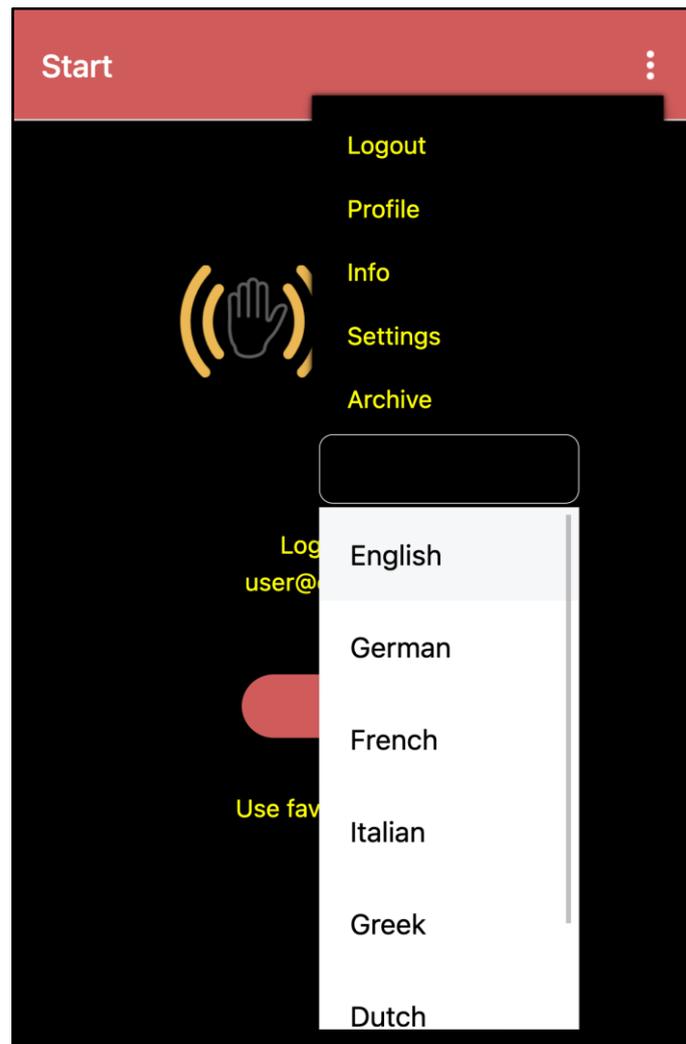


FIGURE 34: LANGUAGE SELECTION MENU IN DROP-DOWN MENU APPEARS IN WHITE WHILE IN DARK MODE

- ➔ When you have enabled the dark mode, the screen flashes briefly white during navigation between the different screens. This is hard on the eyes during navigating in dark mode.

3.2.5 Wording issues (across languages)

- ➔ “Voice style” should be replaced by “Language Level”
- ➔ Language level should be decided at the same time as the language choice.
- ➔ Language level should also be proposed for the avatar.
- ➔ “Info” should be replaced by “About EASIER”
- ➔ “Speech” under the microphone icon should be replaced by “Voice”
- ➔ “Camera” under the camera icon should be replaced by “Sign Language”
- ➔ The “Recording” setting was not understood by some, who wondered if this was another way of accessing the archive files.

3.2.6 Technical issues

- ➔ The maximum number of characters that can be put into the text input is not working. One participant could enter 2100 of 2000 possible characters.
- ➔ One evaluator observed a difference between the apps on the laptop and the mobile phone. The language options in top right are not there on the mobile phone, but they are on the desktop. It shows an empty window, where the word “English” is missing.
- ➔ Standard settings could not be saved.
- ➔ “Use favourite settings” did not work.
- ➔ Some reported that the recording of videos did not work.
- ➔ The audio recording did not work for all participants.
- ➔ If you go in full-screen mode, the app crashes.
- ➔ When you have to activate the camera, the system asks you to do that twice.
- ➔ It was reported that not all icons were active at all times (e.g. recording of videos)
- ➔ Dummy login did not work for some participants.

3.2.7 Other suggestions

- ➔ Provide access to the settings from all pages. Others suggested a horizontal menu bar available everywhere.
- ➔ Add information on how and for what purposes the EASIER app could be used.
- ➔ Allow the production of several translations one after the other.
- ➔ Create a “Save” button to validate language or profile choices, or have some kind of visual confirmation.
- ➔ At this stage with the click dummy, nothing happens when you try a translation. Starting with text input, nothing happens upon going to the next stage. A message saying “This is just a dummy, translations are not yet available” or something similar would have been nice.

3.3 FEEDBACK ON APP EVALUATION PROTOCOLS

- ➔ **The app is better tested on mobile phones or tablets, instead of on computers.**
Some testers commented that there was too much white space on the desktop version. This may suggest that people are not feeling comfortable using the current design for the app user interface on the web browser, and it should be discussed in the consortium if it is needed to release a different design for computer use. One user who is familiar with the latest technologies pointed out that the interface is not “web responsive”, causing the white space issues.
- ➔ **Participants found it difficult to first play with the app and then comment on it afterwards.** They would prefer to immediately give feedback while working with the app.
 - The DGS facilitator changed the task to have two separate groups of discussion and asked participants to take notes while discussing. Afterwards, they presented their collected feedback to each other. This was very useful as the notes taken by the participants could be used for the analysis of the evaluations.
 - The LIS facilitator asked the participants to create a document in which they write down the things they observed during the testing, then asked one participant to start with observations, starting with their notes, and then asked the other participants to speak on a point made when they also had observations on that point, this was to avoid repetition and redundancy. Once the first participant's remarks were finished, the second participant moved on, but with this method the topics already discussed were not dealt with again. In this way, the observations were shared fluidly, at the same time containing the different points of view of all participants.
- ➔ **Participants also reported that the overall goal and use of the app were not clear to them** (a possible explanation is that the missing translation module in the app hindered the evaluation of the full potential of the app).

4 TRANSLATION

The current version of the translation models was tested in the evaluation. These models were used to generate ambidirectional translations to be evaluated by participants. Thus, participants saw and evaluated translations of utterances both from signed to spoken language and vice versa. Spoken language was represented by written text, and sign language was represented by linguistic glosses, as these modes are used as the input and output of the current translation systems (see Deliverable 4.2 **Translation models** for more details on the translation models).

The first translation models are only available for evaluation for two of the project's seven language pairings: DGS-German and BSL-English. This is due to data availability, as these translation models are based on existing available language resources including the DGS corpus and BSL corpus. The translation component was therefore only tested by the DGS and BSL groups. Evaluations were carried out by partners at IDGS and DCAL, the institutions that host the respective sign language corpora. The DGS and BSL facilitators from these institutions were familiar with linguistic glosses and tried to recruit participants who also had some experience reading glosses. Particularly for DGS, several participants were recruited who had some experience within the DGS corpus project.

In this task, model-generated translation was presented alongside the original source material for participant evaluation. The procedure was slightly different among the two language groups. For the BSL groups, participants were presented with the source videos from the BSL corpus, alongside linguistic glosses, when evaluating BSL > EN translation. This was because it was difficult to find participants who were confident enough reading glosses to evaluate the translation solely on the basis of glosses as input. For the DGS groups, participants were presented with only linguistic glosses when evaluating DGS > DE translation. This is because the DGS facilitators were able to find and recruit enough participants who were confident working with glosses as input. For both language groups, participants were presented with only glosses for translations from spoken to signed language (EN > BSL and DE > DGS).

In the following section, we present first the overall impression of participants, followed by the evaluation of each direction sign > speech, then speech > sign. Then we report feedback on the evaluation procedure, questions collected by the facilitator and finally we summarise recommendations.

4.1 OVERALL IMPRESSION

There was a mix of positive and negative reactions to the translation component. On one hand, participants were clear that it needs a lot of work before going mainstream. They all acknowledged clear issues that needed to be dealt with before the translation could reach an adequate standard for users to adopt. While some were pessimistic, calling the MT 'rubbish', others were generous in acknowledging the difficult task of automatic translation and viewed the examples presented as progress.

In fact, participants from all groups were positively surprised with how well the machine translation performed despite omissions and mistakes. In fact, for one translation from DGS to German (sentence 4), the DGS hearing group preferred the machine translation (MT) over the human translation (HT) as they judged it to be closer to natural German than the HT which was closer to DGS. Similarly, the DGS deaf group preferred the MT over the HT for

one sentence from German to DGS (sentence 3). In particular, participants noted that the machine did much better with shorter utterances, despite struggling with longer ones.

4.2 SIGNED TO SPOKEN LANGUAGE

TABLE 2: DGS>DE TRANSLATION SENTENCES

DGS > DE				
	Input	Machine translation	Human translation	Issue/Comment
1	JUNGE6 ICH2	Ich war noch jung.	Ich bin jung.	Verb tense, adds new information (noch)
2	NETT1 EINDRUCK1 STRAHLUNG2	Er war nett und hatte eine Strahlung.	Der macht schon einen ganz netten Eindruck.	Pronoun gender
3	\$NUM-HUNDERTER1 METER1 SPRINGEN2 STARK3	Der 100-Meter-Lauf war stark.	100-Meter-Lauf und Weitsprung, darin ist er richtig stark.	Missing context for MT 'He'
4	DASSELBE2 FORMAT1 ERLEBNIS1 ICH1 DU1 DASSELBE2	Dasselbe habe ich auch erlebt.	Ich habe das auf die gleiche Weise erlebt.	Prefer MT over HT
5	ICH1 TAUB-GEHÖRLOS1 WARUM10	Ich fragte mich, warum ich gehörlos bin.	Warum war ich gehörlos?	Verb tense
6	KEIN5 PROBLEM2 FAX1 FERTIG1	Das ist kein Problem, sondern ein Fax.	Das ist gar kein Problem, es gibt ja schließlich auch Fax.	
7	\$GEST-ABWINKEN1 NÄCHSTE1 TAG1 \$INDEX1 ZEITUNG1 PRESSE1	Am nächsten Tag habe ich in der Zeitung gelesen.	Am nächsten Tag haben die Zeitungen alle davon berichtet.	Omits gestural information, adds new information (gelesen)
8	\$INDEX1 SCHÖN1 FAHRRAD1	Das ist schön, mit dem Fahrrad zu fahren.	Ist aber schön zum Fahrrad fahren.	
9	ARZT1 WOLLEN2 NUR2 GELD1 VERDIENEN1 \$GEST	Der Arzt will nur Geld verdienen.	Den Ärzten geht es nur ums Geldverdienen.	Very good MT, original more sign supported speech
10	SICH-VERSTEHEN1 HÖREND1 WIMMELN1 TRAINING1	Wir haben uns gut unterhalten und haben viel trainiert.	Ich konnte gut mit den Hörenden zusammen trainieren.	
11	ICH1 FROH1 GLÜCK2 \$INDEX1 ELTERN1 ABLEHNEN2	Ich war froh, dass meine Eltern abgelehnt haben.	Ich bin froh und habe Glück gehabt, dass sie es abgelehnt haben.	
12	DU1 SCHWIMMEN1 HIN-UND-HER1 ANSTRENGEND5	Wenn du zum Schwimmen bist, ist das anstrengend.	Wenn du hin- und herschwimmst, das ist doch unheimlich	

	SEHR7		anstrengend.	
13	\$PROD \$PROD VERWIRRT-SEIN1 SCHON1 ICH2 ERSTARRT1 HILFLOS1 \$GEST	Ich war völlig verwirrt, weil ich mich so verunsichert habe.	Dort war ich dann vollkommen verwirrt, ich war unsicher.	
14	BEDEUTUNG1 TAUB- GEHÖRLOS1 HÖREND1 INTEGRATION1 WAS1	Das bedeutet, dass die Gehörlosen integriert werden.	Das bedeutet, dass Gehörlose und Hörende zusammen dahin gehen, oder wie?	
15	UND5 \$GEST HÖREND1 ICH1 GEFÜHL3 UNTERHALTUNG1 VERGANGENHEIT1 KAUM1 SELTEN2 BEREICH1	Ich habe das Gefühl, dass ich mich kaum unterhalten habe.	Und auch mit Hörenden unterhielt ich mich nur selten.	
16	ABEND2 IMMER1 FÜHREN1 \$INDEX1 ERKLÄREN1 WAS1 MORGEN1 MACHEN1 \$INDEX1 DANN1	Am Abend haben wir uns immer erklärt, was wir morgen machen sollten.	Am Abend hat der Führer immer erklärt was wir am nächsten Tag machen würden.	
17	ICH1 HIER1 SCHULE1 \$NUM-TEEN1 \$NUM- ZEHNER1 \$NUM-ORD1 SCHULE1 \$GEST-OFF	Ich war in der Schule 1989 in der Schule.	Ich bin also 1950 in Köln in die Schule gegangen, das war meine erste Schule.	Numbers inaccessible to MT
18	ICH2 TRINKEN-FEIN1 ZUSAMMEN3 KAFFEE2 \$GEST-ÜBERLEGEN1 TRINKEN-FEIN1 \$GEST-ÜBERLEGEN1 BECHER2 \$GEST-OFF ICH2 SCHOCK2 ERSTARRT1 ICH2 KANN1 LAUFEN8 \$GEST-OFF	Wir haben Kaffee getrunken und ich war schockiert.	Wir saßen am Tisch mit Kuchen und Kaffee und da war ich so geschockt, dass ich nicht mehr herumlaufen konnte.	

TABLE 3: BSL>EN TRANSLATION SENTENCES

BSL > EN				
	Input	Machine translation	Human translation	Issue/Comment
1	EMBARRASSED G:DISMISS	I was embarrassed.	It was awfully embarrassing, oh well.	
2	MONTH FOUR02 WEEK-FOUR02 DEPTH	It was four weeks ago.	It lasted for a month, the whole four weeks.	
3	LEAVE SCHOOL PT:PRO1SG FROM-TO LONDON	I left school in London.	After I had left school I had moved to London.	Wrong verb (come to London)

4	TWO INTERPRETER DSEW(1-VERT)-AT: HUMAN DSEW(1- VERT)-AT:HUMAN ONE MAN WOMAN DSEW(1-VERT)- AT:HUMAN	It was an interpreter for two years.	There were two interpreters, one male and one female at the side.	Missing "male", "female", adding "years"
5	PT:POSS1SG BROTHER PT:POSS1SG SISTER02 PT:PRO3PL ALL HEARING SHOULD SPEECH SOME SHOULD	My brother and my sister were hearing.	Because my brother and sister are hearing, they thought we should all speak.	Missing information

For this direction of translation, participants from both languages preferred the HT over the MT, as they felt the former contained less mistakes and omissions than the latter. The DGS deaf group pointed out, however, that the machine did best with the one sentence (sentence 9) that was more sign supported speech than true DGS.

Participants mentioned that **missing context made it difficult to evaluate the sentences**. For DGS, deaf participants also remarked that it was not always 100% clear to them the meaning of the sign language input utterance. In fact, they themselves would translate the 'human translation' of several sentences including 3, 4 and 8, differently given that they do not have access to the original video and surrounding conversation but instead simply the glosses out of context. This is because the human translation included extra contextually implied information. For example, in sentence 4, the MT misses the human referent 'He' which is not included in the gloss but only in the HT, likely because 'He' was already the topic of conversation.

In addition to instances of subject drop in sentence 4, participants also brought up **concerns about other elements of sign language that are underspecified lexically but derived from context**. For example, DGS deaf participants note that pointing (glossed as \$INDEX for DGS or PT for BSL) can have a range of possible meanings that would have to be previously established in discourse and would be unavailable to the MT. Similarly hearing DGS participants noted that information about time/verb tense is also often established in discourse (or else using spatial relationships between signs which are also unavailable to the MT via glossing).

Across groups, participants commented that the **HT was not always necessarily correct**. The BSL deaf group also pointed out for sentence 2, that the human translation is wrong and includes more content than present in the source. Hearing DGS participants also mentioned that just based on glosses, sentence 1 was mistranslated by both the machine and the human; instead, they would translate "Ich bin ein Junge". This evaluation may also be related to the verb tense of the input, MT and HT: both DGS groups noted that it is not clear how the machine gets tense information, given that tense is not encoded in the glosses.

Generally, participants found that **the machine translation omitted information**. BSL participants noted that the MT missed important and obvious content from the glosses, such as the inclusion of MAN and WOMAN in sentence 4. DGS participants from both groups also commented that additional 'non-linguistic' glosses such as \$GEST which mediate the signer's attitude, are missed in the MT's German text (for example in sentence 7).

Furthermore, the DGS hearing group pointed out some instances in which **the machine overinterpreted the input**, adding elements to the MT that were not in the original

utterance. An example of this is in sentence 1, where the MT adds ‘noch’ which changes the meaning of the utterance, or in sentence 7, where the MT adds ‘gelesen’. The same group also were curious about pronoun choice when going from DGS where pronouns are not gender marked, to German where pronouns have obligatory gender marking, for example in sentence 2.

Participants also pointed out **clear mistakes or mistranslations in the MT**. DGS hearing participants were puzzled when the machine misinterpreted what seemed like clear input from the signed utterance in sentence 3: instead of translating 100-metre jump, the MT produced 100-metre run from the gloss \$NUM-HUNDERTER1 METER1 SPRINGEN2 (HUNDRED METER JUMP). DGS deaf participants brought up the strange MT output in sentence 2, in which the sign STRAHLUNG2 was translated as Strahlung (radiation). This appeared to be a clear mismatch with the HT of a person creating a positive impression. Here participants pointed out that the MT was slightly off in selecting the correct German word to convey the proper intended meaning of the DGS utterance.

DGS participants also brought up the case of signs with multiple possible translations into speech, which are under metalinguistic discussion within the signing community with respect to political correctness. This conversation arose based on the translation of the sign TAUB-GEHÖRLOS in sentence 14 to German “gehörlos”. Participants suggested that the app offer multiple translation equivalents, in this case, both “taub” and “gehörlos”, or offer users the opportunity to select and save their preferred term.

Participants from both language groups noted that **by relying solely on glosses, the machine translation missed critical information** not captured in those glosses with respect to non-manuals, mouthing, sign directionality, prosody. The BSL hearing group touched on issues identifying clausal boundaries and spatial relationships between subjects and objects, to which they attributed mistranslations in sentences 3-5.

Furthermore, DGS deaf signers highlighted that **translations (in both directions) missed emphasis conveyed for example, by intonation and speed**; by relying only on written form it is impossible to know where to place emphasis. DGS hearing signers, who were familiar with the glossing system of the corpus also pointed out that because the model does not take into account the notation used for numbers, it is impossible to correctly translate sentences with specific numbers, such as sentence 17.

4.3 SPOKEN TO SIGNED LANGUAGE

TABLE 4: DGS>DE TRANSLATION SENTENCES

DE > DGS				
	Input	Machine translation	Human translation	Issue/Comment
1	Eine gute Frage.	GUT1 FRAGE1	GUT1 FRAGE1 \$INDEX1	
2	Kennst du das?	KENNEN1 DU1	KENNEN1 \$INDEX1	Non manuals
3	Das Baby war gehörlos.	BABY1 TAUB- GEHÖRLOS1	\$INDEX1 TAUB- GEHÖRLOS1 \$GEST- OFF	Prefer MT over HT Verb tense
4	Der Mann war sofort	MANN1 SOFORT2	MANN7 TOD2	Aspect incorporation

	tot.	TOD2		
5	Du gebärdest so für OPA?	DU1 OPA6 DU1	OPA4 DU1	Lexical variants
6	Ich wohne da in der Nähe.	ICH1 WOHNUNG1 NAHE1 \$INDEX1	ICH1 WOHNUNG1 NAHE2	
7	Kein einziger Gehörloser hat dort gearbeitet.	\$NUM-EINER1 EINZIG1 TAUB- GEHÖRLOS1 ARBEITEN1 \$INDEX1	KEIN3 TAUB- GEHÖRLOS1 BEREICH1	Negation omitted
8	Ich muss dir etwas wichtiges sagen.	WICHTIG1 MUSS1 SAGEN1 WICHTIG1	ICH1 BESCHIED1 WICHTIG1 WAS1	
9	Und dann gab es natürlich viele Streitereien.	UND-DANN2 STREITEN1 \$GEST- OFF	\$INDEX1 KONFLIKT1 \$GEST-OFF	
10	Es wäre schön gewesen mit Dolmetschern.	SCHÖN1 DOLMETSCHER1 \$GEST-OFF	SCHÖN1 MIT1 DOLMETSCHER1 EINSTELLEN-ARBEIT1 \$GEST-OFF	
11	Wir haben auch später viel darüber gesprochen.	SPÄTER10 SPRECHEN1	SPÄTER10 ERZÄHLEN4	
12	Und auch mit Hörenden unterhielt ich mich nur selten.	HÖREND1 GEBÄRDEN1 ICH1 SELTEN2 \$GEST- OFF	UND5 \$GEST HÖREND1 ICH1 GEFÜHL3 UNTERHALTUNG1 VERGANGENHEIT1 KAUM1 SELTEN2 BEREICH1	
13	Der macht schon einen ganz netten Eindruck.	\$INDEX1 GUT1 NETT1 DRUCKEN- PAPIER1 \$GEST- OFF	NETT1 EINDRUCK1 STRAHLUNG2	
14	Kinder zu haben, ist für Gehörlose schon anstrengender.	KIND2 ANSTRENGEND1 FÜR1 TAUB- GEHÖRLOS1 \$GEST-OFF	KIND2 TAUB- GEHÖRLOS1 MEHR1 AKTIV3 ANSTRENGEND1	
15	Am nächsten Tag haben die Zeitungen alle davon berichtet.	MORGEN1 ZEITUNG1 \$GEST- OFF	\$GEST-ABWINKEN1 NÄCHSTE1 TAG1 \$INDEX1 ZEITUNG1 PRESSE1	
16	Immer wieder wurde ich gestört, das war nervig.	ICH1 STÖRUNG1 \$GEST-OFF	STÖRUNG1 \$GEST- WÜTEND1 \$GEST-OFF	
17	Wir konnten den Mauerfall ja vorher nicht riechen.	\$INDEX1 MAUERFALL2 KANN1 RIECHEN2	MAUERFALL1 WIR2 RIECHEN3 NICHT1	

18	Viele junge Familien leben gerne in Hamburg in der Stadt.	VIEL1 JUNG1 LEBEN1 GERN1 HAMBURG1 STADT2 \$INDEX1	VIEL1 FAMILIE1 JUNG1 FAMILIE1 GERN1 IN1 HAMBURG1 STADT2 WOHNUNG2 FAMILIE1	
19	Ich bin froh, dass ich heim kann und zurück an meinen Arbeitsplatz.	ICH1 FROH1 NACH- HAUSE1 ARBEITEN2 PLATZ9	FROH1 HEIM6 ARBEITEN1 PLATZ9	

TABLE 5: EN>BSL TRANSLATION SENTENCES

EN > BSL				
	Input	Machine translation	Human translation	Issue/Comment
1	What for?	WHAT	FS:FOR WHAT	MT needs mouthing ("what for")
2	I don't know.	KNOW-NOT	KNOW-NOT PT:PRO1SG	
3	My parents had gone out.	PT:POSS1SG PARENTS GO-TO	GO	Missing context (very short HT)
4	My Dad and I were walking there together.	PT:POSS1SG FATHER WALK- AROUND	WITH FATHER PT:PRO1PL-TWO DSEW(BENT2-HORI)- MOVE:HUMAN	Missing content (I)
5	It was awfully embarrassing, oh well.	EMBARRASSED	EMBARRASSED G:DISMISS	Missing "oh well"
6	I thought, "What is that?"	PT:PRO3SG WHAT	WHAT NOTICE PT:PRO3SG	
7	It was good, I improved.	PT:PRO1SG IMPROVE	IMPROVE	Missing content (GOOD)
8	What's the wall like?	WALL WHAT	PT:DET WALL	

In general, **participants seemed more pleased with the machine translation output for this direction**. In one case (sentence 3), DGS deaf participants even preferred the MT over the HT, as the former was more specific than the latter. Deaf BSL participants similarly evaluated out of the 8 examples, three were correct (sentences 1-3), two were acceptable (sentences 5 and 6) and three were either completely wrong, ungrammatical, or missing important information (sentences 4, 7 and 8).

The DGS hearing group complimented the inclusion of the palm-up gesture, glossed \$GEST-OFF, in several translations, noting that it made the translations seem more natural (sentences 9,10, 12-15). The BSL deaf group on the other hand commented that for sentence 5, "oh well" could have been better translated into sign. Here participants suggested using OH-WELL, a palm-up gesture glossed in the BSL corpus as G:WELL.

As mentioned for sign-to-speech translation, **participants commented on the lack of critical non-manual information** that was not represented in sign glosses. For example, deaf DGS participants pointed out that in sentence 2, non-manuals would be critical to

evaluating the translation as correct, as they are necessary for question formation. Similarly, BSL deaf signers highlighted that for sentence 1, the machine translation would be correct, if the manual sign was accompanied by appropriate mouthing.

Relatedly, the DGS hearing group discussed non-manual negation with respect to sentence 7. While negation is included in the original signed utterance (KEIN3) and the German text (Kein), the MT does not pick up on the negation. Participants discussed that given that non-manual negation is not represented in glosses but is represented in human translations, it may 'confuse' the MT in the training data, and cause it to omit negation in this sentence. Indeed, the participants noted that the glossed signs are in principle okay but would need to be accompanied by non-manual negation marking to accurately translate the sentence.

The BSL deaf group also commented on clear omissions in the MT output, for example in sentence 4 where the "I" (as in "My father and I") is missing, or in sentence 7, where the sign GOOD is missing.

The DGS hearing group also brought up **verb tense for speech to sign**, when discussing sentence 3. Since the German text is specified for past tense "war gerhörlos", but the DGS is not specified for tense, they discussed whether or not the MT should translate tense and how to do so going into DGS. They suggest one way of faithfully translating past tense might be using the sign GEWESEN1 (acknowledging that this use of GEWESEN1 subject to regional variation but common in Hamburg).

Furthermore, the both DGS groups identified the **difficulties for the machine in selecting among sign variants** for a particular translation. Sentence 5 in particular highlighted how this issue becomes critical in metalinguistic discussions about sign choice. Participants also highlighted instances in which signers may want to choose a sign that is in line with their personal preferences. For example, DGS participants brought up the sign FRAU: aside from considerable regional variation in signs for FRAU, the most common variant glossed FRAU1⁴ is consciously rejected by some DGS signers as being sexist. Participants pointed out that should the MT use the most frequent as default (i.e., FRAU1), users should be able to adjust and select variants they are comfortable with.

On the other hand, the DGS groups discussed that **multiple glosses could be used for the same sign and the same sign can have multiple meanings**, an issue that arose when discussing sentence 2. The hearing group noted that while \$INDEX1 and DU1 are both used in the DGS corpus, \$INDEX1 is more commonly used from their experience as annotators. Indeed, in the original corpus glosses for sentence 2, \$INDEX1, while the MT output gives DU1. The hearing group also discussed that since the gloss \$INDEX1 is underspecified for direction, it may cause confusion in the MT output. Both groups note that the sign \$INDEX1 can also have a wealth of different meanings depending on context and direction.

Participants also were puzzled by the lack of context when comparing HT and MT. For example, the BSL hearing group discussed sentence 3, noting the extremely short original signed utterance, compared to the relatively detailed English text. They speculated that the machine worked on a sentence-by-sentence basis, with no surrounding context, while the human translations were excerpts from longer transcriptions.

For DGS and BSL groups, facilitators noted that there was **significantly less discussion of examples for this translation direction**. While they suggested this may be due to the fact that it was the last task after a long day, or that it was simply an uncontroversial task, it is

⁴ FRAU1: https://www.sign-lang.uni-hamburg.de/meinedgs/types/type16458_de.html#type6295

also possible that the design of the task may not have been sufficient to stimulate conversation given the lack of signed translation output (see following section).

4.4 EVALUATION PROCEDURE/METHOD

With respect to the method of evaluation, both participants and facilitators alike commented that **lack of context made the translations difficult to understand and evaluate**. Given that utterances are taken from corpus annotations, human translations may contain more specific detail than available from the glossed utterance because annotators can draw from the context of the surrounding conversation. An example of this is seen in the DGS > DE example 3, where the referent is pronominalised in the gloss but specified in the human translation. Indeed, DGS participants reported that for some utterances they would suggest a different human translation, given the input glosses they were presented with.

Relatedly, participants in the BSL hearing group, composed mostly of academic researchers, questioned whether or not it was a **'fair test' to evaluate the machine translation on the basis of corpus data**. They pointed out that it is common for people to modify their language production when interacting with/via machines (e.g., 'phone voice': more clear enunciation when speaking on the phone), and this input may be more easily translated for the machine than conversational corpus data, which is intentionally collected to be as naturalistic as possible. This suggestion may be particularly pertinent given the issues the MT has with discourse phenomenon such as subject dropping, pointing to refer back to referents and establishing time/tense via context as described in section 4.2, as these are highly common in discourse but may be less common in signing directed towards a translation app.

In addition to this, the BSL deaf facilitator noted that **sentences were too abstract and felt random**, suggesting that more practical examples from potential use cases may be more useful to evaluate (than utterances pulled from the corpus, embedded in the context of random conversation).

Both the BSL and DGS hearing groups also made clear that **it was extremely difficult to evaluate translation output simply via glosses**, and would need to see signed avatar output to properly judge accuracy. As the facilitator for the BSL hearing group put it, without a signed translation output, "it becomes a strange hypothetical exercise where the participants have to guess what the translation output is like and then attempt to evaluate that imagined output".

Particularly for the DGS hearing group, who had experience as annotators on the DGS corpus, the reported discussion of the speech>sign was often quite meta-level and sometimes was not clear if it was directly applicable to the translation evaluation. For example, one participant noted that they would rate the translation quality of speech>sign differently if she saw the original input videos compared to the avatar signing, which seemed a bit beside the point because this means that you would judge the MT on how well it would recreate the original based on human glossing and human translation. One solution to reduce confusion might be to create novel sentences and have both the machine and a human signer translate them so there is a clear progression. Another solution may be clearly labelling each section so that it is clear that the 'input' is back translated text, and the 'HT' is the original signing.

Finally, a note on timing. While the BSL groups made it through all the sentences, the DGS groups had a considerably longer list and did not manage to discuss all sentences within the allotted time frame of one hour.

4.5 QUESTIONS

Participants had several questions and concerns that came up throughout the testing that facilitators collected within their reports:

- Will participants be able to annotate their own data or not?
- Will the machine be able to translate one-handed signing?
- How long does it take for the machine to translate 1 sentence?
- How will the machine catch up to new signs/changing signs?
- How will the machine deal with sign language classifiers?
- How will the machine deal with idiomatic expressions when translating, e.g. English to DGS?
- Will the MT take into account production speed? For example, if an utterance is spoken quickly, will the signed output be signed quickly as well?
- Will regional variation be taken into account?
- Does handedness throw off the MT?

4.6 RECOMMENDATIONS

Aside from the discussion above, participants had a few explicit recommendations. Regarding method, there was a clear wish to see signed translation output from an avatar, as it was difficult to evaluate the translation quality solely based on glosses. Furthermore, participants and facilitators wanted to have a more real-world feel to the task, by including utterances that were more in line with the projected use-cases of the app in day-to-day settings. It was also noted by the DGS facilitators that participants did not fully understand how MT works, and they suggested that the task should begin with more background information on MT, with the facilitators prepared to answer general clarification questions.

For the final product, participants expressed a desire for text output alongside speech output to be able to check the reliability of the translation models. They also wanted the possibility of post-editing, to be able to adjust the output of the translation models to ensure the correct intended meaning is expressed, offering a new translation themselves if necessary. Participants also suggested having a disclaimer in the app about translation accuracy, to make it clear that translations may not be 100% correct. DGS participants also recommended that the app give users the opportunity to select among translation equivalents, both when going from speech to sign (e.g., the case of FRAU) or from sign to speech (e.g., the case of TAUB-GEHÖRLOS). They did however acknowledge that this feature would be time-consuming and would only be useful when the output language is known to the user.

Participants had some recommendations for the translation itself. One participant recommended to improve model subtlety, one could feed the model with several signed sentences with subtle variations and the same written sentence. Other recommendations

were language-specific, particularly for DGS. Participants suggested that for translating speech to sign, in order to convey past tense, the sign GEWESEN1 could be used. They also recommended the use of the person agreement marker (PAM) in DGS (glossed as AUF) to better translate agreement from speech to sign.

Finally, participants had recommendations about tailoring the translations to their preferences. They suggested that the app could present multiple translation choices, and remember users' lexical variant choices (similar to the online translation tool DeepL). They also suggested that the app could provide an option to choose a user's preferred regiolect and use this setting to automatically select signs that fit the region.

4.7 CONCLUSIONS

Several noticeable themes emerged across both translation directions. Participants seemed concerned about the ability of the MT to pick up on important 'paralinguistic' elements such as speed, prosody and gesture. They also were generally concerned about the MT adding random elements, dropping obvious things and mistranslating the intended meaning of the original utterances. Relatedly, participants were deeply concerned about subtlety of expression, ensuring flexibility in the MT for word/sign selection, idioms, and changing language. Participants also noted that the MT has trouble adapting to things that are lexically underspecified in sign languages but are obligatory to encode in corresponding spoken languages, such as gender and tense. Many other critiques stem from the inadequacy of glosses to represent sign language, which is a rising issue in sign language linguistics (see for example discussion by Hochgesang, 2022). Several issues concerned important aspects of a sign language that are traditionally not encoded in glossing but absolutely critical to grammar and understanding, such as non-manual markers.

Participants had much fewer concrete recommendations than in the app or avatar section. This may be due to the fact that participants did not fully understand how machine translation worked. The translation models are somewhat of a 'black box' where input goes in and output comes out, and given that no participants were MT experts this may have limited the concrete recommendations (and overall feedback) given on this task.

Nevertheless, it is also clear that the background of participants affected participant attitudes and quality of feedback. Those groups with some experience in sign language research, such as the DGS groups, gave more linguistically oriented feedback and were more generous, while groups with less of a linguistics background such as the BSL deaf group took translations at face value and were therefore more strict with their critique. While both types of feedback are useful user input, perhaps they may be better separated out into different formats. For example, smaller focus groups of experts could give deep linguistic feedback and large-scale layperson evaluation can provide an overview of acceptability using more structured questionnaires or rating tasks.

5 AVATAR

The avatar evaluation was presented in the form of a questionnaire, which was followed by a group discussion. This questionnaire was developed by ATHENA to collect feedback on the avatar from deaf community members (see Deliverable 2.1 **Interim sign language avatar** for more information about the development of the questionnaire). Appendix B contains screenshots from the questionnaire.

The questionnaire was prepared for the four sign languages for which the avatar is currently available: GSL, DGS, DSGS and French Sign Language. Thus, eight groups (deaf and hearing from each of the four languages) completed this component of the evaluation.

First, participants were given the link and directed to complete the questionnaire.⁵ The questionnaire contained instructions filmed in each of the relevant sign languages. The questionnaire aimed to collect feedback on the acceptability and legibility of the avatar, by presenting sample videos of the avatar and asking participants to rate them. At the end of the questionnaire, participants were invited to upload a video of their feedback. After completing the questionnaire, participants returned to the group to discuss the avatar.

In the following section, we report on both the structured feedback from the questionnaire, as well as the qualitative feedback that emerged in the following focus group discussion. This qualitative feedback adds nuance to the questionnaire results, with detailed discussion of the issues identified by users, and bringing valuable insight into the criteria by which users judge the sample sentences. We also briefly mention feedback from the app section of the evaluation, in which some groups saw the avatar settings as part of the app interface.

5.1 QUESTIONNAIRE RESULTS

The questionnaire first asked for background demographic information about participants, with respect to their age, gender, sign language proficiency and language learning context. Participants then viewed and rated videos of the avatar. First, the participants viewed single signs and were asked to rate “How well does she sign”, using a five-point scale, ranging from “Very well” to “Bad”. Individual signs varied from language to language. Participants then proceeded to judge the avatar signing 5 complete utterances. Utterances had the same semantic content across all languages, and corresponded to the following English sentences:

1. Hello, I’m ready to begin.
2. Could you repeat that?
3. Sorry, I didn’t understand.
4. Please wait, response is pending.
5. Thank you for using our service. Bye!

Participants were shown each utterance in three stages, and in each stage, a video of the avatar signing an utterance was presented in combination with a rating question. In stage 1, participants were asked to rate “Did you understand what Paula signed?”. In stage 2) participants were asked “Did Paula sign like a human?”. In stage 3) participants were presented with a video of a human signer alongside the avatar signing the same utterance,

⁵ The questionnaire can be found at: <http://sign.ilsp.gr/slt-eval/>.

and were asked “Did both of them sign the same thing?”. Images from the questionnaire can be found in Appendix B.

5.1.1 Single signs

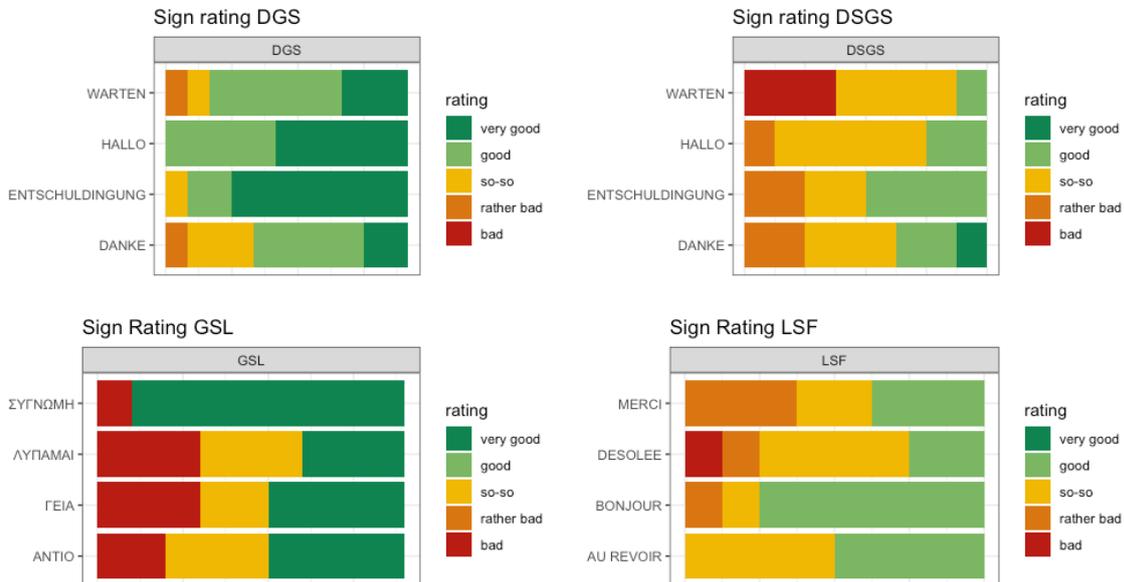


FIGURE 34: PARTICIPANT RATINGS ACROSS LANGUAGES FOR INDIVIDUAL SIGNS (TOP LEFT DGS, TOP RIGHT DSGS, BOTTOM LEFT GSL, BOTTOM RIGHT LSF)

GSL signers show strong extremes with the same signs being rated on opposite sides of the spectrum. DGS signs are mostly rated as good or very good, with no signs rated as outright bad. LSF and DSGS are mostly so-so or good with few signs rated as very good.

5.1.2 Multi-sign utterances

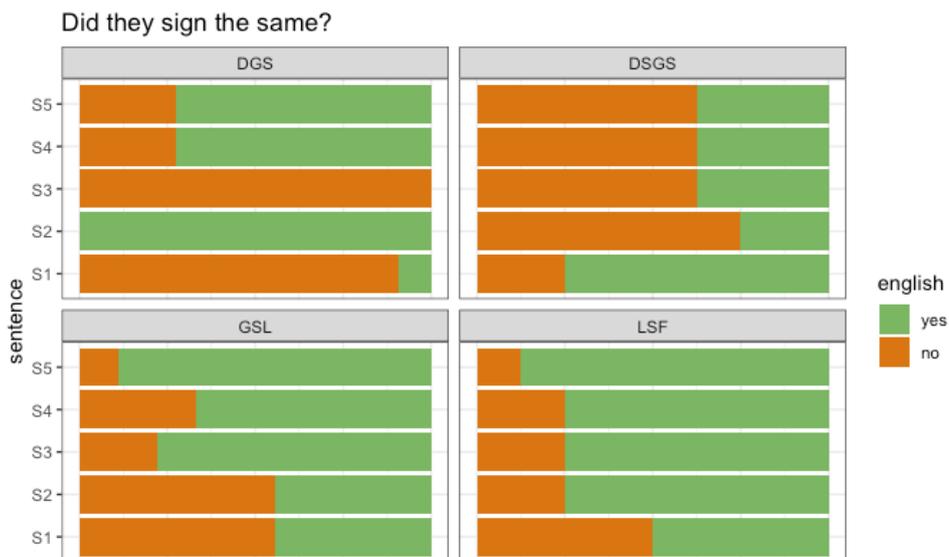


FIGURE 35: PARTICIPANT RATINGS ACROSS LANGUAGES FOR “DID THEY SIGN THE SAME?”

When comparing Paula to a real signer, the responses were mixed. For LSF, participants generally had positive responses to this question, with the lowest scoring item, sentence 1

receiving 50% yes and 50% no votes. For DGS participants were more divided by sentence. Sentence 2 was unanimously voted to be signed like the human yet sentence 3 was unanimously voted to be different. Sentence 1 also received overwhelmingly negative responses, while sentences 4 and 5 were generally positively judged to be similar to the real human. DSGS was also mixed but for all but one sentence (sentence 1), the majority of participants judged that Paula did not sign the same as the real human.

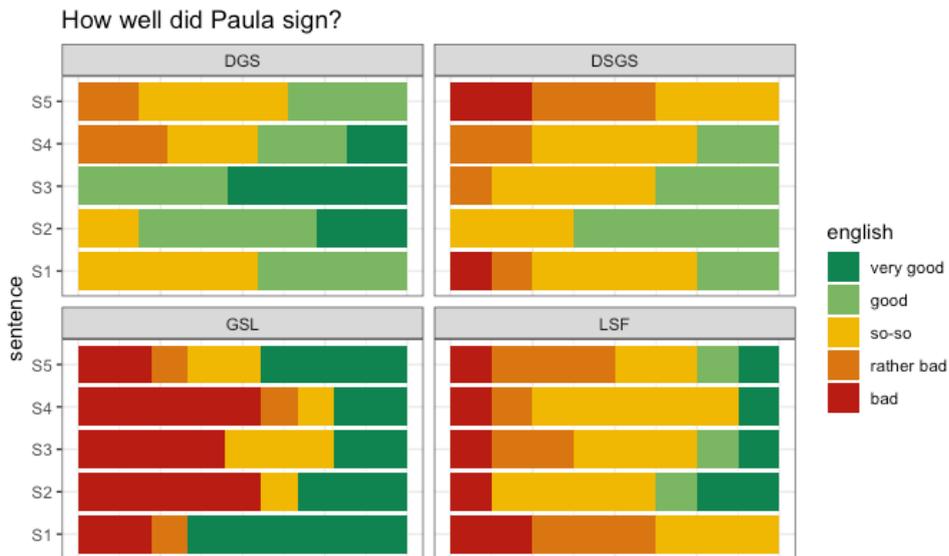


FIGURE 36: PARTICIPANT RATINGS ACROSS LANGUAGES FOR “HOW WELL DID PAULA SIGN?”

In response to the question of “How well did Paula sign”, participants were again highly varied by group. While some sentences such as DGS sentence 3 received generally positive evaluations, others such as LSF sentence 1 and DSGS sentence 5 were evaluated as between so-so and bad.

Like in the single sign judgements, GSL signers showed strong extremes in both questions. DGS signers again judge most things as so-so or better. DSGS ratings are somewhat lower and LSF shows a range of responses including very good and bad for the same sentence. (In LSF, one participant responded bad to both questions for all the sentences).

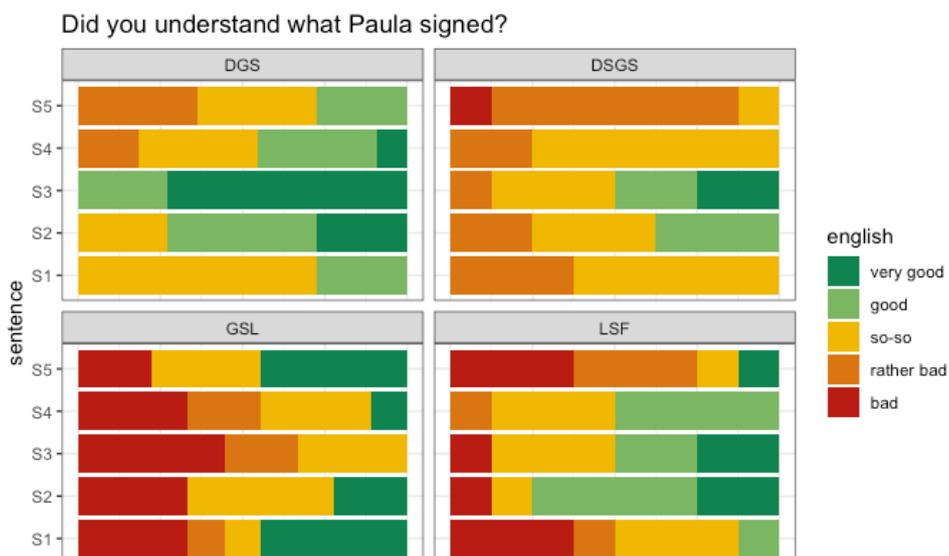


FIGURE 37: PARTICIPANT RATINGS ACROSS LANGUAGES FOR “DID YOU UNDERSTAND WHAT PAULA SIGNED?”

Responses for “Did you understand what Paula signed?” were relatively similar to “How well did Paula sign?”, with the same sentences receiving generally positive (e.g. DGS sentence 3) and generally negative responses (e.g. DSGS sentence 5).

Given the distribution of responses for each sentence, it may be useful to combine a rating with another task/question that allows participants to pinpoint exactly what they don’t like in the sentence. For example, it may be possible that while participant A can overlook a poorly formed sign and still judge the sentence overall as good, participant B may judge the sentence as bad due to the same error. While the focus groups dealt with Paula’s signing in general, this would allow participants to give feedback specifically on particular sentences.

5.2 FOCUS GROUP DISCUSSION

5.2.1 Overall appearance and legibility

With regards to Paula’s appearance, participants were highly detail-oriented, one group even complimented her highly realistic fingernails. Some features of her appearance received both positive and negative feedback, however. While a participant from one (hearing) group complimented the general appearance and colour contrast between skin and clothing, another (deaf) participant recommended that darker clothing would better serve deafblind users, for example, those with Usher’s Syndrome. The shadow cast by Paula was noted as positive by one hearing group (they did note how important it was to have the shadows precisely match the body), however one hearing interpreter in another group did not like it as they felt that interpreters and translators are typically lit to have no visible shadows. The logo on Paula’s clothes was viewed positively by one (hearing) group, but as unnecessary by another (deaf) group who suggested the logo could be located in the background, off the avatar’s body.⁶ The same deaf group also made the point that the current avatar is white, but it would be important to have it available in different ‘skins’ that better represent people of colour in Europe.

Particularly those who have had experience with other avatars received Paula positively. For example, the DGS groups who had experience with Anna⁷, another signing avatar, commented that Paula was much better than other available signing avatars. Responses from deaf LSF participants revealed an interesting age effect. On one hand, young people who have experience with video game avatars were more accepting of Paula. They pointed out that while she was not perfect, she was a work in progress and would inevitably improve much like video game avatars have across the years. On the other hand, older participants in the same group were more sceptical about the avatar and less willing to accept it at this stage of development, noting “*Everything was difficult to understand for me, I’m not happy with this solution.*” (LSF_D4).

Across the board, participants were clear that more work must be done to make Paula’s signing more human-like. Several groups labelled Paula as ‘robotic’. This critique stemmed from several issues, but primarily due to the lack of non-manual movements of the face, head and body (see next subsection). Another issue brought up was the quality of movement. One group remarked that it appeared as if Paula only could make use of the x, y and z axes of motion sequentially, but could not combine them - meaning she could move

⁶ The latter feedback came from the LIS deaf group, who did not evaluate the avatar in its entirety but had feedback based on the avatar video they saw in the app.

⁷ The avatar Anna was developed in the projects ViSiCAST, eSIGN and Dicta-Sign.

her hands forward then right but not make a smoothly combined forward-rightward motion. However, all feedback was not critical. Participants also complimented those aspects of Paula's signing that they felt were highly realistic. For example, the same group that critiqued the motion path also complimented the joint motion, noting that when signing HELLO (in DGS) the waving motion comes from the wrist, not the elbow or shoulder, making it look more natural.

Overall, the attitudes towards Paula can be summed up as *better, but still not human-like*. Following this, one group reflected in detail whether the goal was to produce an avatar that was as human as possible, or to produce an avatar that is as legible/understandable as possible.

5.2.2 Non-manuals

It was clear the standards for Paula were very high. Some participants watched and re-watched videos in detail, looking for minute muscle movements, particularly in non-manuals. They were attuned not just to linguistic movements, but also non-manuals that conveyed emotion, and small details such as blinking that all contributed to overall human feel. It was generally agreed that Paula needed more non-manuals: existing movements needed to be amplified and additional movements needed to be added.

Face: linguistic and non-linguistic

Participants were generous with complimenting what they felt were improvements in Paula's non-manuals. They noticed and remarked upon small facial movements such as frowns, blinks, lip and eyebrow movements as being extremely positive additions. However, they noted several aspects of the face needed work. As one participant put it "*It's not bad, the avatar has evolved. It has become more refined, but the work on facial expressions is not finished. The avatar's face is still empty.*" (LSF_D4)

In general, participants felt Paula's signing missed an affective component, and the lack of emotion contributed to her '*cold*' or '*robotic*' demeanour. In the example sentences, several groups pointed out that when signing THANK-YOU, HELLO or GOODBYE, a small smile was missing from Paula's face. The lack of a smile evoked strong negative reactions in participants, "*She says goodbye with a cold face, it really bothers me*" (LSF_D4). One group also noted that the shape and position of Paula's eyes and eyebrows "*gave an intonation to the speech, as if she was always in a state of surprise*" (LSF_H2).

In her current state, Paula's non-manuals are too reduced and not easily perceivable at first pass, especially when viewed on a small screen, e.g., a smartphone. One deaf participant noted, "*Yes, there are expressions. The frown, when I take a detailed look, I see it. But when I keep an overview, I don't see the facial expressions; I have to look carefully to see them.*" (LSF_D3). As one deaf group explained, it is less tiring to look at an avatar when the facial expressions are properly amplified and easy to pick up on, and contributes positively to the overall understanding of the signed message.

Related to this, participants wanted to see more comprehensive facial expressions. For example, one group pointed out that a smile should also be accompanied by a change in eyebrow position and eye aperture to reflect the natural muscle movements of a human face.

With regards to the eyes, participants mentioned issues with both eye aperture and eye gaze. Following the discussion of amplifying non-manuals, participants from multiple groups remarked that both opening and squinting of Paula's eyes should be more intense to be

more human-like. Furthermore, participants pointed out a lack of coordination of eye gaze when referring back to established indexes in space: Paula simply stares ahead, and does not change eye gaze direction while signing. Indeed, in sign languages, coordinating eye gaze when establishing and referring back to entities is a critical part of the grammatical system (akin to verb agreement in spoken languages).⁸ In fact, the same group that noted issues in gaze changes also pointed out that whenever Paula established an index in space, it was unclear who or what that index referred to.

Another issue brought up by several groups was that some signs were missing non-manual components, meaning they were incorrectly formed phonologically.⁹ For example, GSL signers pointed out that the sign NOT-UNDERSTAND was not understandable as it lacked the appropriate non-manuals. Similarly, LSF signers pointed out that some signs required non-manuals to distinguish them but those non-manuals were not present. (These comments also applied to non-manuals of the mouth and body).

Mouth: mouthings and mouth gestures

Participants noticed Paula's mouth movements as a positive, but had a lot of critiques on how this could be further improved and refined.

First, similar to mentioned for the face, participants wanted to see more comprehensive movements to accompany mouthings. They pointed out that mouthings should also be accompanied by changes in cheek position to look more natural.

Next, LSF groups identified that for some signs, such as HELLO and THANK YOU, Paula used English mouthing to accompany the LSF sign. This was met with some degree of outrage, and it was reiterated that the mouth must match the surrounding spoken language of the sign language, in line with typical mouthing practices.

DGS participants from the hearing group pointed further mismatches between Paula's mouthing and typical mouthing practices. First, they mention that in DGS like in other sign languages, typically only the word stem is mouthed,¹⁰ instead of the entire word, as Paula does. Second, they point out that the timing is not right, and Paula's mouthings begin and end too early.

Furthermore, participants noted that the mouthings were not clear enough, and stressed the importance of increasing their prominence. As one deaf participant noted *"I grew up speaking. I became a signer later on. So I also need to lip-read, but there, on the avatar's face, there was nothing, it was empty"* (LSF_D3). In some cases, when there was no mouthing, it prevented comprehension of the whole utterance. For example, in GSL, the sign SERVICE can be used to mean both *service* and *employee*, however as the sign was not accompanied by mouthing, participants did not know which meaning to take from the manual sign alone.

In addition to mouthings, mouth gestures were also discussed. DGS signers noted that the mouth gestures were also not always accurate, for example, the sign BALD1A (see figure below).

⁸ See for example, Garcia & Sallandre (2020)

⁹ Non-manuals are critical aspects of SL phonology, see for example, Pezendich (2020)

¹⁰ See for example, Bank et al (2011)



FIGURE 38: WRONG MOUTH GESTURE FOR DGS SIGN BALD1A

Body

Multiple groups complimented the inclusion of shoulder and body movement, noting that it made Paula look more human. However, other participants from the DGS groups noted that there was not enough movement of the upper torso, and in particular left to right movements of the torso were 'choppy'.

5.2.3 Prosody

Participants across groups commented that Paula's signing style was hard to understand, particularly due to the speed of signing, the transition between individual signs and the lack of prosodic markers.

Several groups noted that there was no change in speed to Paula's signing, which was uniformly very slow. One participant mentioned "*the rhythm is slow; I looked for a speed-up button on the videos!*" (LSF_H1). Her slow signing generated misunderstanding among participants, and contributed to the 'robotic' look.

Multiple groups pointed out that the transition between individual signs was too slow. One deaf LSF participant explained that it drew attention to unimportant movements (a sentiment which others in the group agreed with), "*All the gestures are broken down, we see all the movements. It's misleading because my eyes are drawn to movements that I don't usually see. [...] It's a question of rhythm. The pace is slow and it gives significance to things that are not important.*" (LSF_D4). For example, in the sign sequence pictured below, deaf LSF signing participants were confused by the impression of an intermediary sign in frame (2) between (1) the signs UNDERSTAND and (3) NOT.



FIGURE 39: FALSE INTERMEDIARY SIGN BETWEEN UNDERSTAND (1) AND NOT (3) (LSF)

In addition to transition between individual signs in general, participants pointed out that transitions both within and between phrases needed improvement. Between signs within

phrases, the transition was at times too slow; for example, BITTE1a WARTEN3 in DGS. In other instances, the transition between phrases was too abrupt, not incorporating pauses.

Indeed, participants lamented the lack of prosodic markers, such as pauses and body movements, to structure signed utterances. As one LSF participant noted, “*Rhythm plays the same role as punctuation. If there is a comma or a dot, we will mark micro-silences. In sign language, there are moments when we stop for a short time before moving on to the next part. These are rhythmic mechanisms which allow us to make sense of the text.* (LSF_H1). The lack of prosody presented a significant challenge for legibility; some DSGS participants reported only fully understood the meaning of an utterance when they watched the human signer in the comparison task.

Taken together, Paula’s slow signing speed and lack of prosody made utterances difficult to understand at first pass, and participants reported needing 2-3 views of a video to get accustomed to Paula’s signing.

5.2.4 Manual sign formation

Some groups brought up issues with manual sign formation that affected legibility. In particular, participants pointed out issues with signs in which the location was the weak hand or body. GSL signers noted that particularly in signs with L or 5 handshapes, when the thumb made contact with the chest it was difficult to see (for example WAIT in GSL). DSGS participants pointed out that for signs with location on the weak hand or body, while the location was not always accurate, the sign was often understandable.

During the focus group discussions, feedback also came up for specific signs signed by the avatar. The table below summarises the formational inaccuracies brought up for signs across the different languages (for more details, see individual language reports in the annex). Some of the inaccuracies reported made these signs, or else the entire utterance unintelligible. Other inaccuracies left the target signs still interpretable but only with considerable effort.

TABLE 6: SIGN SPECIFIC ERRORS

Trouble sign	Language	Description of Issue
SERVICE2A	DGS	wrong mouthing makes sign unintelligible
BALD1A	DGS	wrong mouth gesture
EINTRETEN1A-\$\$SAM	DGS	too choppy
UNSER1A	DGS	too distal
GRENZE1A-\$\$SAM	DGS	too linear
SERVICE	GSL	can’t disambiguate between meanings without mouthing
NOT-UNDERSTAND	GSL	needs non-manuals for legibility
SORRY	GSL	wrong handshape and place of articulation
SIGN	GSL	wrong handshape and place of articulation

GOODBYE	GSL	missing a smile
THANK YOU	GSL	missing a smile
WAIT	GSL	occlusion of hand making contact with chest makes it difficult to understand
ANTWORTEN	DSGS	place of articulation: higher in space than expected
VIELEN DANK	DSGS	no change of cheeks when mouth moves
WAIT	LSF	lacking repeated motion
SORRY	LSF	irregular orientation and point of contact, missing shrug
HELLO	LSF	English mouthing
THANK-YOU	LSF	English mouthing
BONJOUR	LSF	missing small smile
PRÊT	LSF	missing semi-circular and descending movement of the active hand and 90-degree rotation of supporting hand

5.2.5 Methodological feedback

Participants and facilitators had feedback with regards to the method of testing, and specific suggestions for improvements in the questionnaire.

At least one group noted that in general the instructions for the task were clear and very good. Several groups, however, reported issues with the video recording feature at the end of the questionnaire and were not sure if this feature worked or not. Some participants reported only seeing a blank screen while for others it seemed to work.

Multiple groups noted issues with the language background section. The hearing LSF group commented that available responses to the question “Where did you learn sign language” did not correspond to users' real experiences. For the hearing DGS group, it was not clear where to find “School”, in the case of learning DGS at school/university. In the same DGS group, the facilitator also found it useful to explain the proficiency levels in ‘Gut’, ‘Mittelmäßig’, etc. in terms of the Common European Framework of Reference (CEFR) for language proficiency.

In terms of the questions themselves, a few feedback points came up. One DSGS group reported that the video format of the human signer was distorted (see Figure 40). Participants from the DGS group were puzzled that one item comparing Paula’s signing to the human signer came up twice (“Sorry, I didn’t understand”), and wondered whether it was simply an error or some form of experimental control.

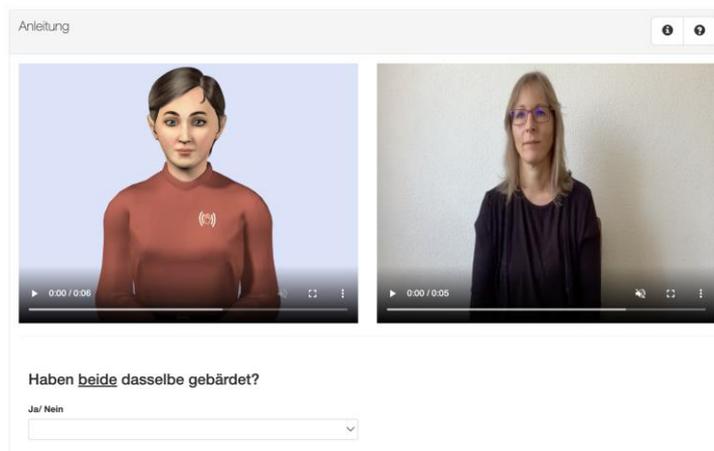


FIGURE 40: DISTORTED IMAGE OF HUMAN SIGNER (DSGS)

The same DGS group discussed the questionnaire format in detail. They questioned whether there was any meaningful difference between the question “How well does Paula sign?”, versus “Does Paula sign like a human?”, as participants would automatically compare Paula to a human standard for the former (this point was also raised by the deaf LSF signers). They also found difficulty with the question “Did both of them sign the same” as they found it unclear whether it referred to the same content or the exact same meaning. From the DSGS group, one participant suggested adding another intermediate response option “Only part of the sentence was understood”.

5.2.6 Recommendations

TABLE 7: AVATAR RECOMMENDATIONS

Theme	Issue	Recommendation	Groups
Overall appearance and legibility	Contrast not sufficient for deafblind users	Make clothing darker to increase contrast with skin tone	GSL
	Inaccurately formed signs	Users and national deaf organisations validate sign accuracy	LSF
	Hard to see handshapes with thumb extended touching chest	Option to rotate signer to get around occlusion	GSL, DGS
Non-manuals	Inaccurate language mouthing	Correct the language of mouthings (LSF-French)	LSF
	No changes of the wider face to accompany mouth movements	Add complexity to facial expressions, e.g., for smile add squint and eyebrow movements	LSF, DSGS
	Mouthing entire word, wrong timing of mouthing	Mouth only word stem, use research to inform mouthing	DGS

		animation	
	Non manuals are too small, hard to see on phone and impede comprehension	Amplify non manuals, add option to zoom in	LSF, DGS
Prosody	Transition between signs is too slow, especially within phrases, giving a robotic look	Smooth and speed up transition between signs	LSF, GSL, DGS
	No sentence level prosody	Add prosodic pauses	GSL, DSGS, LSF
	Slowness of signing generated misunderstanding	Create speed setting for avatar	LSF, DSGS
Method	Existing responses not sufficient	Add another response option "Only part of sentence was understood"	DSGS
	Confusion in language background section	Add CEFR levels, simplify specifying having learnt to sign at university, match learning scenarios to users' experiences	LSF, DGS
	Several participants felt survey missed return button	Add return button on questionnaire	DSGS
	Duplicated question for "Sorry, I didn't understand"	Remove duplicated question	DGS

5.2.7 Other feedback

In addition to those groups who explicitly evaluated the avatar through the questionnaire and focus group discussion, some other groups noticed the image of the avatar available in the settings of the app. This image, along with the available settings in the app to customise the avatar, generated some discussion and feedback that can be found in section 3.1.11 Representation of diversity.

5.2.8 Conclusions

The evaluation of the avatar benefitted from combining a more structured questionnaire with in-depth focus group discussion. From the questionnaire, the range of responses often included both extremes of the rating scale. This showed that participants may have very different standards of what is acceptable, or may interact with the rating scale in very different ways. On its own, the questionnaire is too coarse a measure to understand the specific issues that lead participants to rate a sentence as *bad*, however, including a follow-

up question about where exactly the issue lies in the sentence may be useful. More data from participants in the second evaluation round will be useful to identify larger trends.

In the focus group discussions, however, participants gave strong and clear recommendations about how to improve legibility. Two major themes cut across the groups. First, they require more, clearer and amplified non-manuals. This includes more and amplified mouthings, emotion embedded in certain signs and phonologically correct non-manuals. Next, there must be improvement in Paula's signing speed. At the moment, her prosody is awkward and generates misunderstandings. Participants need to see grammatical use of pauses/body movements to segment utterances and more natural sign transitions.

Other than feedback on how to improve Paula's legibility, this interim evaluation also produced good feedback for improving our methods in the final cycle with several suggestions for improving the questionnaire. The discussions also revealed interesting insights about the target populations, particularly that young people and those familiar with video games may be more accepting of Paula as they are familiar with avatars and the rapid development of this technology.

6 OTHER FEEDBACK

6.1 USE CASES

The focus of this evaluation was to give feedback on the intermediate versions of the individual components. However, on several occasions, while evaluating the components, participants brought up feedback relevant to use case scenarios.

Some participants of the app evaluations named different use cases where they might use the app, such as situations in public transport or while shopping, like asking where one can find goods in a store.

Discussions of the translation component also touched on accuracy and use cases. While it was clear that the current degree of accuracy was not appropriate for professional contexts, both BSL groups discussed that it might be useful for low-stakes basic everyday interactions.

Nevertheless, it should be made clear the degree of translation accuracy is considerably below 100%, one deaf BSL participant suggested a disclaimer in the app to make users aware of this.

7 CONCLUSIONS

Taken together, the results from the interim evaluation can give the technical partners in the EASIER consortium a clear map towards improvement, producing a great deal of specific feedback. Indeed, they demonstrate the importance of integrating end-users into the development cycle and bringing intermediate/prototype systems to the users for feedback.

Nevertheless, some components received more feedback than others. While the app and the avatar received highly detailed recommendations, feedback on the translation component was considerably less. This may be because participants are familiar with apps for translation and have a clear idea of what works well and what they like. Similarly, participants are familiar with natural sign language use, so are able to give clear critiques of where the avatar's signing fails to meet natural, human standards. However, with the translation, participants had many questions about the technology and not much specific feedback. One facilitator suggested that the translation section would be better accompanied by a more in-depth background about how the technology works, to allow participants to better understand what they are seeing.

There was also an overall push to match the evaluation to more real-life scenarios that reflect the intended use cases of the systems. For example, some recommended using more real-life translation sentences than those from the corpus, both as it may give the MT a better chance of increased performance, and it may give participants a better idea of what is acceptable output. Others recommended testing the app in a real exchange between two people to evaluate how cumbersome the interface is. The current navigation in the app required to make a translation and choosing the settings were strongly rejected, as participants felt these back-and-forth steps heavily impaired the process for a smooth communication flow between interlocutors. Evaluators made several comparisons with well-known machine translation tools, such as Google Translate and DeepL, as the interfaces thereof should be taken as a reference for the EASIER project.

Participants also were concerned that the conditions of day-to-day use match their needs, such as one-handed signing, working despite visually 'busy' backgrounds. This feedback should be taken where possible into the final evaluation cycle, as testing in more realistic scenarios can both help users give more targeted feedback and also help users understand and give feedback on intended use cases.

Within the focus group discussions, one theme shone through both the avatar and translation components: the issue of non-manual components of a sign language. The discussions drove home the point that non-manuals are a critical component of sign language grammar and leaving them out of the avatar or machine translation components dramatically affects understanding. Non-manuals do not just serve to add extralinguistic information such as emotion, but they also are integrated into the language structure on all linguistic levels. Without integrating non-manuals in the input and output, we run the risk that the real meaning of signing is unintelligible.

Another important theme that emerged from all discussions was that deaf communities have strong values that need to be reflected in any service designed for them. First, there is a strong commitment to diversity. Deaf groups consistently brought up issues of diversity and inclusion, with concerns that deafblind users, nonbinary users and ethnic minority users feel comfortable and supported using the app. Relatedly, deaf groups were clear about the need for full accessibility for various groups: for example, they pointed out potential issues for 'grassroots' deaf users with low literacy, and the need for contrast/colour-blind-friendly colour schemes. They also are advocates for clear communication about the limitations of the technology, for example advocating for a disclaimer about translation accuracy. These recommendations should all be taken seriously in technical development, and even for those recommendations that are

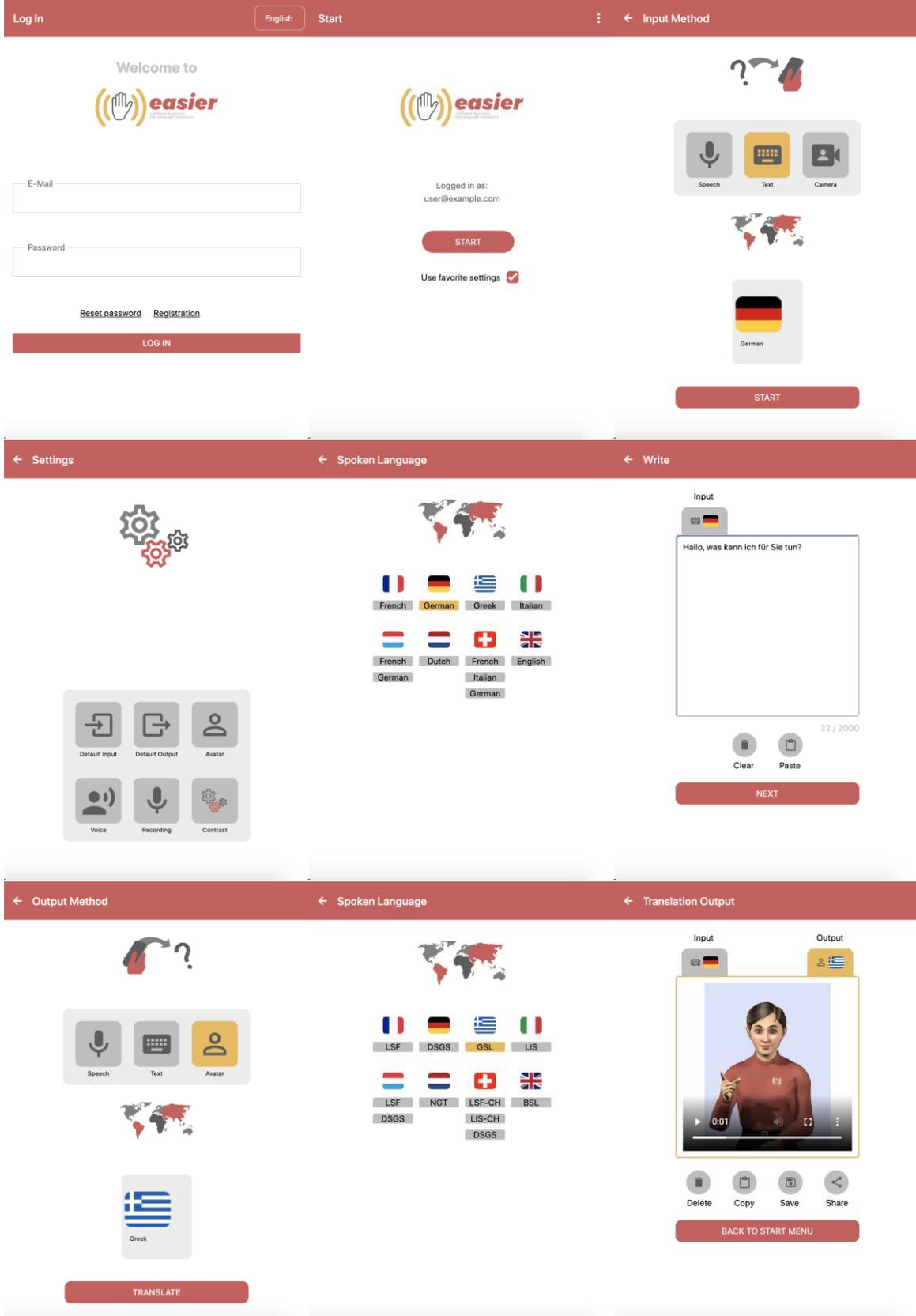
impossible to implement in the life cycle of this project, should be taken into account for future similar projects, as they are likely to affect whether or not deaf end-users adopt the technology.

Finally, it is worth noting that across the board, participants were all happy to be involved and many asked to be contacted for follow-up research. These connections can be further exploited in the final evaluation cycle, where a major recruitment drive is needed. This enthusiasm for involvement indicates a positive degree of interest in the project and further drives home the importance of end-user evaluations.

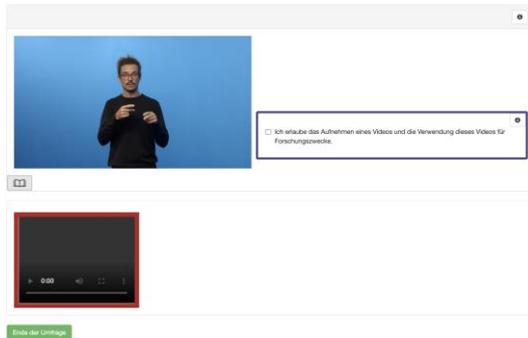
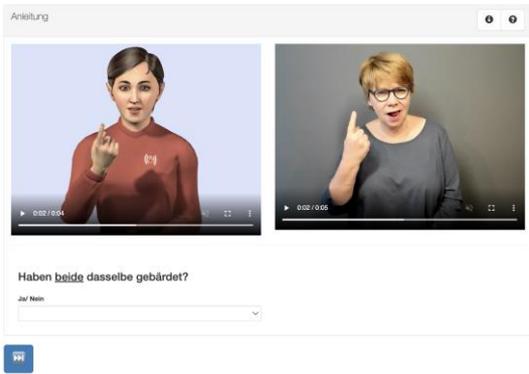
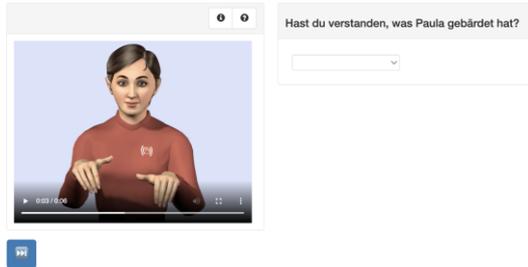
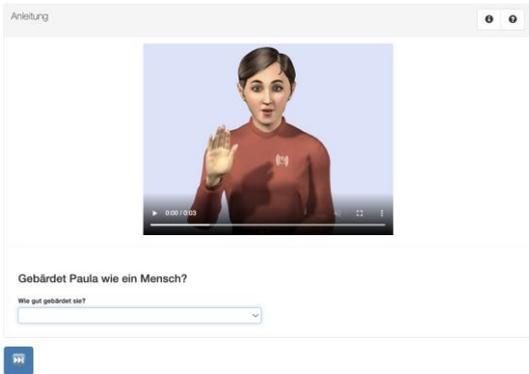
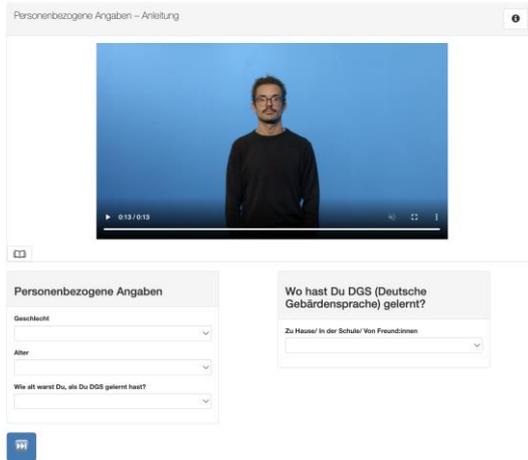
REFERENCES

- [1] Bank, R., Crasborn, O. A., & Van Hout, R. (2011). Variation in mouth actions with manual signs in Sign Language of the Netherlands (NGT). *Sign Language & Linguistics*, 14(2), 248-270.
- [2] Garcia, B., & Sallandre, M. A. (2020). Contribution of the semiological approach to deixis–anaphora in sign language: the key role of eye-gaze. *Frontiers in Psychology*, 11, 583763.
- [3] Hochgesang, J. (2019). Tyranny of Glossing Revisited: Reconsidering Representational Practices of Signed Languages Via Best Practices of Data Citation. Talk presented at the workshop Doing Reproducible and Rigorous Science with Deaf Children, Deaf Communities, and Sign Languages: Challenges and Opportunities. Berlin. September 23, 2019.
- [4] Napier, J., Skinner, R., Adam, R., Stone, C., Pratt, S. and Obasi, C. (2021) A demographic snapshot of the profession: The 2021 census of sign language translators & interpreters in the UK. Chester: Association of Sign Language Interpreters.
- [5] Pendzich, N. K. (2020). *Lexical nonmanuals in German Sign Language: Empirical studies and theoretical implications* (Vol. 13). Walter de Gruyter GmbH & Co KG.

APPENDIX A - APP PROTOTYPE



APPENDIX B – AVATAR QUESTIONNAIRE



APPENDIX C - FACILITATOR REPORTS

BRITISH SIGN LANGUAGE DEAF GROUP	71
BRITISH SIGN LANGUAGE HEARING GROUP	78
GERMAN SIGN LANGUAGE DEAF GROUP	85
GERMAN SIGN LANGUAGE HEARING GROUP	106
SWISS GERMAN SIGN LANGUAGE DEAF GROUP	134
SWISS GERMAN SIGN LANGUAGE HEARING GROUP	139
FRENCH SIGN LANGUAGE DEAF GROUP	ERROR! BOOKMARK NOT DEFINED.
FRENCH SIGN LANGUAGE HEARING GROUP	165
DUTCH SIGN LANGUAGE DEAF GROUP	183
DUTCH SIGN LANGUAGE HEARING GROUP	189
GREEK SIGN LANGUAGE DEAF GROUP	195
GREEK SIGN LANGUAGE HEARING GROUP	205
ITALIAN SIGN LANGUAGE DEAF GROUP	214
ITALIAN SIGN LANGUAGE HEARING GROUP	219



BRITISH SIGN LANGUAGE DEAF GROUP

- **NAME OF PERSON WRITING REPORT: NEIL FOX**
- **DATE OF EVALUATION: 29TH OCTOBER 2022**
- **SIGN LANGUAGE: BRITISH SIGN LANGUAGE**
- **GROUP (DEAF/HEARING): DEAF**

1. METHOD AND PARTICIPANTS

1.1. Facilitator

This evaluation was facilitated by Neil Fox. Neil is deaf, a native BSL user. He is based in DCAL as a researcher and is a member of the EASIER project.

1.2. Participants

There were 4 participants in total, recruited by all recruited by the facilitator. They were recruited by facilitator's contacts. Their age ranged between 18-65, Three were females and one male. Their occupations varied from a BSL teacher, service manager, primary school teacher and a student. There were three native BSL users, and one acquired BSL at early age. No one else was present in the session

1.3. Procedure

The process consisted of two halves divided into sub-tasks:

- Evaluating the prototype web interface
 - Setting up default settings
 - Performing a translation
 - Group discussion / feedback
- Evaluating and discussing example translations
 - BSL to English
 - English to BSL

The entire session lasted 1 hour 45 minutes, with the web interface taking up slightly more than half of that time (just under 1 hour).

1.4. Technical setup

The evaluations were in person. The participants used their own phones, and a laptop for interacting with components. The facilitator had a laptop.

2. FOCUS GROUP DISCUSSION

2.1.App

2.1.1.General feedback

The participants were instructed to find and set up default language settings/preferences and then attempt to perform some translations. Minimal guidance was given, although they were explicitly advised that although they were accessing the prototype via a web interface, they should remember that its primary platform was intended to be as a phone app. They were also advised that translation output was not yet available. Two people used their mobile phones where the other two used laptops.

There were log in issues, all four participants could not log into the app as user@example.com. One of them created an account, and shared it with everyone and they all used that account. The overall responses were generally negative. One participant said it is rubbish. They were not sure to what to do, it was not clear, and navigation is poor.

They had a question about data protection and archive, how long personal information is stored, and 'archive'. They are wondering why archive is there, who maintain this data. Archive might not be the right word but they like the idea of having a fixed, or saved list that the translations they use regularly like favourites.

2.1.2.Theme 1 - Navigation

One clicked on the flag, and text to choose Italian and then nothing came out. She is not sure of the criteria or what is expected task. She wants to look at picture or word and click on it but it is frozen. The steps are not clear, click on flag then spoken language but there are no options or pictures for sign languages.

When you go to default and pick a language for 'text', 'speak' and 'avatar', it seems that you have to have one language for all three and you cannot choose different language for each set ups, it could be useful for travelling to have different languages for the three options.

One said there are too many pages, Google Translate only has one page and very straightforward where this app has too many steps, too many navigations. It wasn't easy to pick settings and there is no guide, demo.

One provided more positive feedback that some elements are good and useful however she knows that the grassroots deaf will struggle with this app. She finds it easy to use, but yes, she agreed that there are too many steps to take and then click on back button, and do it again for next option. It all should be on one page, that the settings should be on the same page.

Examples of comments

translation, input output avatar, lot of forward back forward back.

don't need in and out avatar, all can be on one page

each button has own screen don't need.

settings take while.

One remarked that there is lot of work is needed to create a default setting and this need to be simplified (and deaf friendly). One also expressed concern that situation arise when she arrive

at a location and want to use the app, there is a feeling of too much hassle to do it, it take too long to actually do it, they would revert to default communication set up such as pen paper, gesture, or use Big Word app.

One observed a difference between the laptop and mobile phone app, on the top right, language options are not on app, but on the computer. Where the app shows an empty window, there is a missing word that say English.

2.1.3.Theme 2 – Visual

Some of them were not happy with the symbols used for avatar's gender. One did say they prefer English word – male & female. Some did say that it may not be clear to grassroot people. They weren't sure about 'x' for no gender, and what would non binary look like, they cannot see at the moment.

They had issues with no race options, should have more diversity, size, age, they also expressed concerns about how voice is linked to avatar, please see next section for more details. While choosing avatar, they should be able to choose the voice, or while setting up default, such as language and things, the app can use that information to set up the voice to meet the user's age and characteristics.

One positive thing they liked was the ability to change to dark mode, because this is good for dyslexia and deaf-blind. One person did not know where to find this option, so the other person showed how to find this. Some of them preferred black option. However, they expressed concerns how dyslexic people would find it in the first place because it is very hard to find. It needs to be there in the home page, made clear, and someone suggested an automatic accessibility button immediately on the front page.

They made comments that it felt inappropriate to have a male as vest, and female as t-shirt, this is not universal or consistent, women can wear suit. One person thought it was a vest but did not realise it is meant to be suit. They did not realise it was not gender but for the voice, see next section. This part (including symbols) was not clear to everyone.

They did say that the design is boring, need to make it appealing, very basic, but simplified don't mean boring. look like first day of internet, very basic.

2.1.4.Theme 3 – Voice (Sounds)

One participant expressed concern about avatar's lip patterns although we did not see any avatar.

Issues over accents were mentioned. Why is there a choice between English and American, this is Europe. Where is German accent and so on, why is America there? UK has diverse accents, such as London, Yorkshire et cetera.

How would the app adapt for tonation, to appear interested, aggressive, gentle or polite, there is a toggle for 'on and off' - what does that mean? How would that sound like, as they are deaf, and how can audience understand their voice?

Voice style, the symbols for smart, casual – initially some of the people thought the symbols represented gender but they now realise they represent 'style' but it is not clear however they all agree now. It seems that the younger generation understand the app better when compared to the older participants. They appear more flexible and know how to use the app, so need to consider broad generations (skill set).

One mentioned that they would prefer to have an user ID, when you create your profile, they will then provide an avatar of male/female etc, then they can customise specific features. (Like several default set ups).

They felt that the profile was very basic, as there are only three settings – name, email and id. That is all. There need to be pronouns, how I prefer to be called, what race, age et cetera so people will know what your background and type of person you are.

2.1.5. Recommendations

See above for recommendations

2.2. Translation

2.2.1. General feedback

1. Was the response generally positive, or negative?
2. How was the translation presented to participants?
3. What guidance did the facilitator give to the participants?
4. Which things did participants like, which did they not like?
5. Introduce major themes that came up (can add more than 3 Theme sections)

They were clearer on this task, although it was still not ideal. For the BSL>EN direction, source videos were available for the input BSL as well as glosses. For EN>BSL, only glossed output and written input were available (avatar output was not available). The response were generally negative.

For BSL>EN, first sentence is correct.

Second sentence was wrong, translation is wrong, and the human translation had extra contents, however the human version is better.

Third translation is wrong, it should be left school and COME to London, translation says he left school in London but he moved after school. This should be more like human translation.

Fourth sentence has wrong translation. They were puzzled where the YEAR came from, also no male or female was included. This is extremely easy to identify and they are shocked this is missing, such an important information. They are concerned about the ability of the app to translate sentences; they want the information to be accurate. They feel that sign languages are too difficult for machine to understand, additionally there are regional variations, too many different versions. Machine cannot understand all, including locations of sign.

Fifth sentence, the translation missed half of the information.

ENG>BSL

1 Correct, someone said about mouthing, or different gloss WHAT-FOR

2 correct

3 correct

4 wrong, WALK-AROUND is the wrong verb, the sentence also does not include himself.

5 ok, but can add OH-WELL, but it's fine (however by adding OH-WELL is important to show he is not bothered (context))

6 could add some pointing at the end, (or THINK)

7. missing GOOD. need to sign GOOD

8. like. look not good sentences

2.2.2.Recommendations

General comments were that some translations were okay but in longer sentences there are not enough information being translated. Short sentences seem ok, however when there are two sentences in one there are not information that are included.

BSL teacher added that there is a possibility that conditional sentences should be included in translation such as what if, etc. Topic & comment. BSL grammar.

One person thought the 'machine' was rubbish, extremely behind and providing limited sentences. That we will never get full information, only little information. There need to be a lot more work. Machine would never be 100% it seems to be like a guessing work, BSL interpreters do some guess work, there are never 100% success when translating.

One added that she probably wouldn't use the app, that she would carrying on texting. However, it might be for the younger generation but not for me, it could be appropriate for older generation. If I had a choice between text or sign in the app, I would choose text. The app is suitable for people who cannot text or do not have good English. It is linked to how much accuracy is acceptable for use. It would not be suitable for professional content but maybe everyday basic communication.

BSL signs always change and there are new signs, how will the machine catch up. One questioned how long it takes for machine to translate one sentence. If it is a long sentence, it would be slow to perform? It is good for people with limited written skills, they do need help. However, they do need to bear in mind that it may not be accurate.

It needs a lot of work before it goes mainstream.

(Repeating this paragraph that was included in previous section)

Fourth sentence has wrong translation. They were puzzled where the YEAR came from, also no male or female was included. This is extremely easy to identify and they are shocked this is missing, such an important information. They are concerned about the ability of the app to translate sentences; they want the information to be accurate. They feel that sign languages are too difficult for machine to understand, additionally there are regional variations, too many different versions. Machine cannot understand all, including locations of sign.

3. OTHER FEEDBACK

3.1. Feedback from facilitator (optional)

Both pilot and focus groups provided similar feedback, which were generally negative. They do see the potential of such tool, but I think it was too early in design stage to have feedback, but it is valuable to potentially make lot of design changes and perhaps improve the translation components.

I feel the translation sentences need to have more practical examples, such as every day sentences not random sentences with no context that people are not going to use in their situations. It might provide better translations or examples.



BRITISH SIGN LANGUAGE HEARING GROUP

- **NAME OF PERSON WRITING REPORT: MATT BROWN**
- **DATE OF EVALUATION: 27 OCTOBER 2022**
- **SIGN LANGUAGE: BRITISH SIGN LANGUAGE**
- **GROUP (DEAF/HEARING): HEARING**

1 METHOD AND PARTICIPANTS

1.1 Facilitator

This evaluation was facilitated by Matt Brown. Matt is hearing and is currently based at DCAL and UCL Linguistics as a PhD candidate, has previously been a corpus linguistics research assistant at DCAL, and also currently teaches part-time at the university's Centre for Translation Studies. He is a qualified BSL/EN interpreter but has not been actively practicing in that role since 2018.

1.2 Participants

Five participants took part in the evaluation group. They were all hearing women, ranging in age between their late twenties and early fifties. Four were colleagues at DCAL and had backgrounds in linguistics, psychology or cognitive neuroscience. The fifth was recruited externally and was a practising BSL/EN interpreter. They were each offered a small/nominal sum of money as recognition of their contribution.

1.3 Procedure

The process consisted of two halves divided into sub-tasks:

- Evaluating the prototype web interface
 - Setting up default settings
 - Performing a translation
 - Group discussion / feedback
- Evaluating and discussing example translations
 - BSL to English
 - English to BSL

The entire session lasted 1 hour 45 minutes, with the translations taking up slightly more than half of that time (just under 1 hour).

1.4 Technical setup

The session took place via Zoom, with the facilitator and participants using either their own personal or university equipment.

During the session, it became clear that the use of video sign language examples did not work well over Zoom's screen sharing functionality, with the majority of the participants complaining about the frame rate (especially when video clips were short). This was solved during the session by sending the video materials to the participants by e-mail so that they could access them offline.

2 FOCUS GROUP DISCUSSION

2.1 App

2.1.1 General feedback

The participants were instructed to find and set up default language settings/preferences and then attempt to perform some translations. Minimal guidance was given, although they were explicitly advised that although they were accessing the prototype via a web interface, they should remember that its primary platform was intended to be as a phone app. They were also advised that translation output was not yet available. All of the participants used a desktop computer browser to access the prototype website (Chrome / Edge / Safari).

The response was generally quite critical and tending towards the negative, with a majority (but not all) of them expressing feelings of confusion around navigating the prototype. Their feedback has been summarised here under two main themes.

2.1.2 Theme 1: lack of interface feedback / no feeling of progress

None of the participants felt that the prototype was “intuitive” and two clearly expressed a view that it was unintuitive. A majority of them made remarks during this half of the session about not having much context or guidance. One participant more specifically remarked that they had seen some apps which have a “first time user” experience of some kind, where an outline / overview / tutorial is given and then dismissed, and wondered if something like that would be beneficial.

All of the participants remarked on the strangeness of having to go “back” after selecting options. Two of them said they expected to see some sort of visual feedback after selecting a language and perhaps moving “forward” with a Confirm or Next button – instead, the buttons felt “dead” and there was no feedback to tell you that the option had actually been selected. It took them multiple attempts to realise that the option had been selected and they needed to return to the previous screen by clicking on the back arrow. A couple of the participants remarked on inconsistencies with this approach, noticing a Next button after you select translation languages and go on to enter typed text.

Several of the participants said that they felt that selecting default language inputs was redundant and questioned why it was done, since every attempt to perform a translation seemed to ask you to confirm those language selections again. More than one of the group felt that your last used preference should be remembered and “defaulted to”, and they were not clear about whether they really had to select languages every time they performed a translation or whether they had misunderstood.

2.1.3 Theme 2: visual design

One participant remarked on the strangeness of the size and spacing of interface elements, but felt that this was perhaps due to her technical setup (monitor resolution and browser window size).

One made positive comments about the visual elements of the interface but still felt that without more context it was not always clear what the icons represented.

The majority of the participants found the interface confusing when selecting signed and spoken languages. One noticed inconsistent page headings (clicking on a camera input icon

takes you to a page headed “Spoken Language”). Another expressed confusion about whether the use of national flags was identifying signed or spoken language selection and found the identification of BSL with the Union Jack flag a little confusing. People were uncertain about the meaning of the language input icon that contains small hands, especially when it was used for selected typed or spoken input (they associated hands with signed input).

More than one participant remarked that the world map icon was confusing or misleading due to the highlighted countries not matching the language selections available.

One participant made very positive remarks about the selection of options for the avatar output (gender/clothing colours etc.) but would have liked to have seen them in action.

2.1.4 Recommendations

One participant suggested, and the others all immediately agreed, that there were too many clicks and too much navigation. They envisaged a use of the app out on the streets, having to have a back and forth exchange with someone, possibly passing the phone back and forth; they felt that the app should be striving to make that exchange as easy as possible and minimising the number of screens and clicks/taps, that there was still a lot of work to do here.

One participant had a previous career as a software tester and said that in that role, they would not normally have been asked to evaluate something so early in its development.

One participant remarked on the colour selections and wondered if they were colourblind-safe.

Two participants suggested, and all of the others agreed immediately, that all of the language inputs should be on the same page at the same time, instead of having to back and forth / in and out of several pages of selections. For example, if performing a translation from typed English to BSL avatar output, both input and output languages should be available at the same time as dropdowns or some other element above the text input area.

2.2 Translation

2.2.1 General feedback

In general the participants seemed to find this process a little confusing at times. Feedback was a mixture of positive and negative remarks.

For the BSL>EN direction, source videos were available for the input BSL as well as glosses. For EN>BSL, only glossed output and written input were available (avatar output was not available). Two or three of the participants were able to read glosses to some extent – where participants were finding the glosses opaque or difficult and no video examples were available, the facilitator assisted by giving signed examples of what they thought the glosses most likely represented.

2.2.2 Theme 1: omissions

For the BSL>EN direction, in general the human translation was preferred. It was generally felt that the human omitted less, and made fewer mistakes / better decisions.

One participant felt that the translations were “never” going to account for non-manual information, that they would not perform well where facial expression or motivated use of space or classifier handshapes were key, that (e.g.) clausal boundary markers would not be

identified, or the spatial relationship of subjects/objects would not be identified. They identified this kind of problem with BSL>EN sentences 3, 4 and 5 in particular. The group returned to these kind of discussions/questions on more than one occasion when trying to analyse why some of the machine translations were felt to be inadequate.

The participants had one particularly interesting discussion where they attempted to establish a baseline for “adequacy” depending on the intended use of the app. The consensus seemed to be that since this app is not intended as a replacement for human interpreters or critical situations but more aimed towards shorter “everyday” interactions, a lack of nuance in the translations would probably be quite acceptable. It was felt in general that communicational adequacy could be achieved for some situations even if the translation quality was low, as long as the errors were mostly omissions and not more misleading errors.

2.2.3 Theme 2: other errors

The participants did not have the same feeling about mistakes as they did about omissions. They tended to be much more critical of misidentified signs or intended meanings (for example, BSL>EN sentences 2 and 4) than they were for omissions. In general for the BSL>EN direction it was felt that the machine translations contained many more clear mistakes than the human translations.

For the EN>BSL language direction however, the participants also found some of the human translation decisions quite perplexing. As a group they started to wonder if there was some other context to these examples that they were not privy to and expressed some mild suspicions about the provenance of the examples. They speculated that the machine translations might be working on a sentence by sentence basis (with no “memory” of any surrounding linguistic context) but that the human translations might be sentence-level excerpts from a longer translation; this latter suggestion was prompted by the extremely short human translation in EN>BSL sentence 3.

One of the participants wondered whether the handedness of the BSL users was throwing off the translation quality, asking whether it could identify right-handed and left-handed signers.

One participant made an interesting statement, questioning whether the use of corpus video stimulus was a fair test of the machine translation’s abilities. They made comparisons with hearing people using a special “phone voice” when using voice input (e.g. Apple Siri or Google Assistant) – that people “put on” a special loud/slow/clear tone when they know they are talking to a computer performing speech recognition (because they expect it to go wrong otherwise). It was felt that for signed-to-text translations, the machine might be disadvantaged compared to the human translator if it was not given an analogous kind of “model” signed input.

2.2.4 Theme 3: lack of signed output

The participants generally had much less to say about the EN>BSL direction. In general the facilitator perceived that it was more difficult to discuss a translation into sign language when you cannot see it. In fairness, this was the last section of the evaluation and energy levels may have dropped a little by then.

2.2.5 Theme 4: positive remarks / surprise at the technology

Two of the participants both remarked more than once that they were surprised the machine translations had got so much right, despite the omissions and mistakes.

2.2.6 Recommendations

Most of the participants expressed a clear wish to see the signed translation output, that it was very hard to evaluate quality from glosses and discussion alone. There were no other clear recommendations from the translation evaluations beyond the feedback above.

3 OTHER FEEDBACK

3.1 Feedback from facilitator (optional)

In general this facilitator concurs with the feedback from participants around lack of avatar output, that this half of the translation evaluation for the EN>BSL direction does not work well at all without it – it becomes a strange hypothetical exercise where the participants have to guess what the translation output is like and then attempt to evaluate that imagined output.

I feel like that use of a remote Zoom meeting worked adequately well – it was very difficult to organise five people into the same place and time and Zoom was likely the only way that number could be achieved on this occasion (six could not be recruited at the same time and place even with Zoom). However, the interactions would probably have been smoother, and technical issues fewer, if it had been possible to hold the session face-to-face.



GERMAN SIGN LANGUAGE DEAF GROUP

- **NAME OF PERSON WRITING REPORT: MARIA KOPF**
- **DATE OF EVALUATION: 25 OCTOBER 2022, 10:00–16:00 CEST**
- **SIGN LANGUAGE: GERMAN SIGN LANGUAGE**
- **GROUP (DEAF/HEARING): DEAF/HOH**

1 METHOD AND PARTICIPANTS

1.1 Pilot Study

Two pilot studies were conducted, one for the deaf and hard of hearing (HoH) group, one for the hearing group. This should ensure that both facilitators were prepared for the task. The pilot study for the deaf/HoH group was conducted on 20.09.2022. Present was one deaf participant, the deaf facilitator (Julian) and the organizer (Maria). All parts of the planned study were tested. The pilot study lasted six hours including breaks of approx. 90 Minutes. A summary of both pilots (one for the hearing group, one for the deaf group) can be found in the Annex under 0 Feedback Pilot studies.

Based on these findings the following adjustments were made:

- The app is tested on mobile phones or tablets instead of computers.
- Slides of the Machine Translation (MT) task will be presented stepwise to better navigate the discussion.
- The video shown at the beginning of the Avatar task is moved to the beginning of the meeting and will be skipped in the Avatar task.
- Minor changes in the slides to better support the facilitators.

1.2 Facilitator

Julian Bleicken was the facilitator for the deaf group. He is a research associate at the Institute for German Sign Language and Communication of the Deaf (IDGS) at the University of Hamburg. He works for the DGS Corpus project since 2015 and is a lecturer at the University of Hamburg. The focus of his work in the DGS Corpus project is on the transcription of signed texts. Julian holds an M.A. in Sign Languages of the University of Hamburg. Julian is deaf and learned DGS in the beginning of his teens.

Julian knows the EASIER project as Maria regularly reports on it in internal DGS-Korpus project meetings, but can be seen as external to the project.

1.3 Participants

For recruiting participants, an invitation mail in written German and DGS (see Annex 0) was sent via an internal mailing list at the IDGS, a mailing list of the professional association of sign language interpreters in Northern Germany (Berufsverband der Gebärdensprachdolmetscher/-innen in Norddeutschland (BGN) e.V.), the newsletter of the Deaf Association Hamburg (GLVHH – Gehörlosenverband Hamburg) and to the interest group of Deaf students in Hamburg (iDeas Hamburg). Additionally, contacts to members of the signing community in Hamburg were used to invite people personally.

Six participants were recruited. The ages ranged from 24 to 53 years old. There were five men and one woman. 4 participants are deaf, 2 hard of hearing. Two participants grew up with DGS from the beginning, two started to sign around the age of 4/5 years old and two started in their teens.

The organizer (Maria) was present through breaks and the last half hour of the focus group meeting to answer more specific questions about easier and the MT.

1.4 Procedure

1.1.1.1 Procedure

Both focus groups (deaf/HoH and hearing) tested three components of the EASIER project: the app, the machine translation and the avatar. Both groups followed the same procedure. A set of slides was used to navigate through the sessions (see Annex 0) which were organised as follows:

Welcome and Introduction

- Welcome
- Presentation of the planned time schedule

Consent

- Presentation of consent form, answering of questions regarding consent
- Signature of the consent form by participants

Presentation

- Introduction of participants (name, age and language background)

Background

- Project video: <https://www.youtube.com/watch?v=jmyEUqixIOU>
- Further information on EASIER
- Information on evaluation procedure

Components Test I – App

- Explanation of task
- App tested by participants individually: <https://easier-integration.nuromedia.com/>
- Group discussion on the following questions in two smaller groups:
 - How easy was it to make a translation?
 - How intuitive is the use of the app?
 - How clear is the layout of the app (visually)?
 - Which functions are still missing?
- Discussion within the whole group

Break

Components Test II – MT

- General information on MT
- Explanation of task
- Discussion of MT sentences

Break

Components Test III – Avatar

- Explanation of task
- Online questionnaire filled in individually: <https://sign.ilsp.gr/slt-eval/>
- Group discussion on avatar

Final round of questions and discussion

1.1.1.2 Duration

In total the focus group meeting lasted 6 hours including breaks.

Introduction and Background: 20 minutes

Components Test I: 80 minutes

Components Test II: 75 minutes

Components Test III: 45 minutes

Final round: 35 minutes

Note: Two participants were not present from the start, but 15min and 70min into the first session.

1.5 Technical setup

The evaluation took place in person at the IDGS. Participants sat in a half circle for general input and group discussions (see Figure 41), for the individual tasks they used computers (iMacs) facing the wall provided by the IDGS or their own phones/tablets. A large screen (behind the facilitator) was used to show the slides and task material.



FIGURE 41: HUMAN SETUP

The meeting was recorded from three different angles using three GoPro HERO6 cameras (see **Error! Reference source not found.**). A resolution of 1080p with 25 fps. Two cameras were set to wide angle, one to normal angle. During session II one camera only recorded for 20 minutes and then stopped for unknown reasons.

2 FOCUS GROUP DISCUSSION

2.1 App

2.1.1 General feedback

The facilitator first asked the participants to play with the app and try to make a translation. Later he offered them the above-mentioned questions to be discussed within the group.

The participants used their own smartphones with Android and iOS operating systems and Firefox, Chrome or Safari web browsers.

The general feedback was mixed, tending to be more negative than positive.

In the further discussion the topic of the target group was raised. Participants discussed that there are different groups ranging from children to old people and that the app should meet a certain average to be appropriate for everyone. Problems estimated for old people were the size of the screen, the intensity of mouthings on the avatar and the use of fingerspelling. It was also stated that old people will probably not be the main target group.

2.1.2 Video Input

The participants criticised that the video screen is not mirroring the image of the signer. Placing oneself in the focus of the camera was difficult as it is counter intuitive.

The format of the video screen is too narrow to record the whole signing space. It should be wider, or horizontal not vertical and at least 4:3 or 16:9. The screen format of the app FaceTime was named as a good example to follow.

Other participants preferred a vertical video screen so that they don't have to turn their device by 90 degrees, but the video screen should be bigger with less white space around it.

Participants thought about the case when one person is filming another person who is signing. In this case one should be able to change between front and back camera of the device.

Participants missed the function of a double click to change the video screen into full screen, as the button to change into full screen is very small and was not found by everybody.

One participant had trouble finding the "Record" button. They would prefer the button to be next to the video screen, not inside it.

2.1.3 Navigation through the app

Some participants had trouble navigating through the app. They would like selections being saved by clicking on the button and immediately being put forward, or via a "Continue" button. The need to press "Back" to go along the settings was strongly rejected.

For some participants this was the only part of the app that was not intuitive. They found the structure and setup logical and clear to use. For others, further points of irritation came up when navigating through the app:

- There were too many screens between adding the text and the translation, including "Back" steps. This was considered unintuitive.

- After finishing the first translation, there should be a way to go back to the text and add a new text, instead of being thrown back to the start of the whole process and having to begin from anew. Also, the text should be saved in the input window.
- When saving settings, it was not clear that they were saved as the button did not respond in any manner but making the user go back. This was confusing to the participants. By clicking the “Save” button the app should automatically go back to the screen before.
- Some participants were irritated that they had to click exactly the arrow-button but could not click on the word next to it to navigate back (see Figure 42).



FIGURE 42: LANGUAGE SETTINGS IN APP

Sometimes participants closed the tab (not on purpose) and had to open the website again. Each time they had to enter user name and password again, which they did not appreciate. They would prefer to have this information saved for further uses. Standard settings could not be saved which was confusing to the participants.

One participant strongly rejected the app as it is so different from other translation apps (he named *Google Translate* and *DeepL*). This participant would prefer to have input and output next to each other instead of on different screens with multiple steps/screens in between. As well as having the settings for input and output language on the same screen. For this participant the navigation was way too complicated and the deviation from the common standard was seen as contrary to user-friendliness. In the discussion with the other participants, it was mentioned that next to each other could be a problem on a small screen, but above each other was thought to be a good solution.

Note by the organizer: No participant tested the function of changing the avatar settings. Participants did not find the function of switching between input and output on the translation page (see Figure 43).

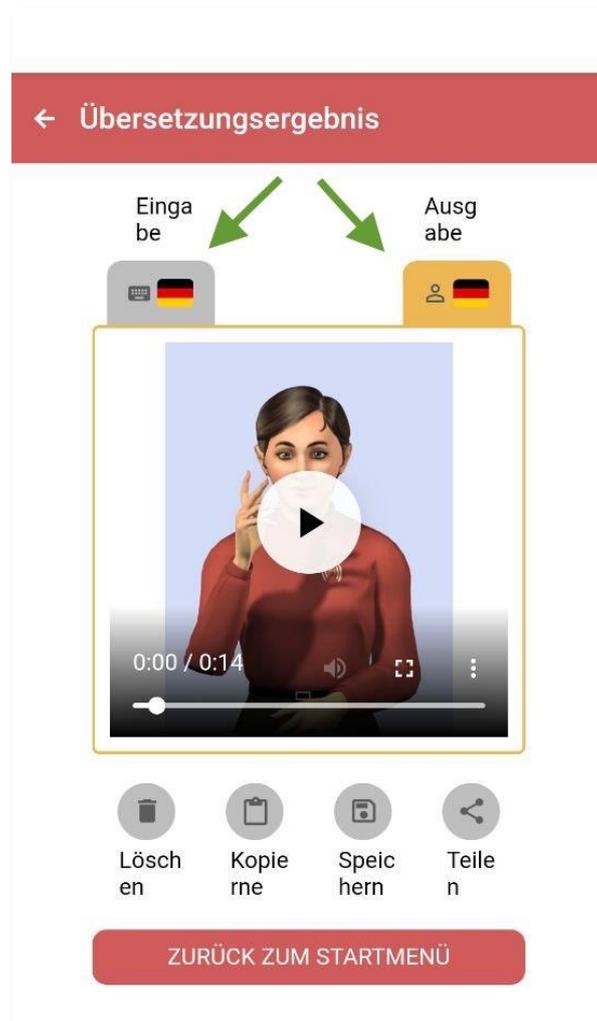


FIGURE 43: TRANSLATION OUTCOME

2.1.4 Language selection

Participants found it confusing to only have spoken languages as input languages and sign languages as output languages.

2.1.5 Dark mode

Participants described that with the dark mode many areas/buttons of the app are not distinct any more. Frames around buttons and areas would be preferred (see Figure 44 vs. Figure 6).



FIGURE 44: LIGHT MODE

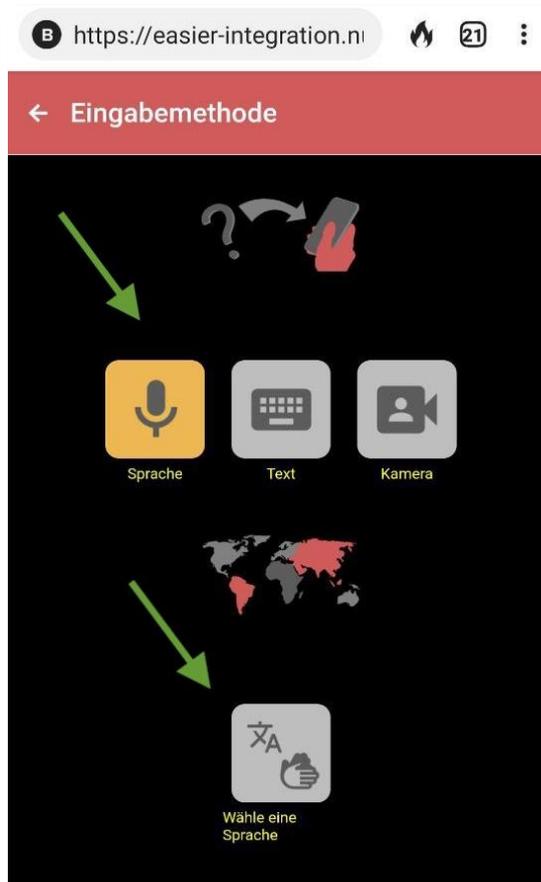


FIGURE 45: DARK MODE

Others stated that the dark mode integrated in the app works fine, but if the device itself is in a dark mode the app is not shown properly, e.g., the upper bar is not distinct from the rest of the screen anymore, errors are missing. This means that the device set up has to be changed to light mode and then the app itself into dark mode which seems rather complicated.

Participants would like the font colour in the dark mode to be white and not yellow or to have the option to choose between the two.

2.1.6 Layout

The general layout was clear, especially with the three modes of input/output. Although one participant criticised the symbol of a camera for the input of signing. For him it was not clear that the camera is the one that means signed input, as a hearing person could film themselves while speaking.

The participants criticised that not all buttons are visible on their screen immediately, but that they had to scroll down to push the “next” and “start” button.

2.1.7 Archive

Files in the Archive should be named by keywords from the content that was translated instead of numbers.

2.1.8 Typos

Participants found typos in the app: “Aktiviere Kammera” should be “Aktiviere Kamera”.

Participants found some English terms in the German version of the app.

2.1.9 App icon

One participant had strong feelings about the icon of the app. The participant felt that it is not noticeable enough, which would make it hard to find the app in the list of all apps on one’s device. The participant stated that there definitely should be two hands.

Another participant likes the EASIER logo as it is, and another one stressed that the EASIER app is not only for signing but also for written and spoken texts so the logo should stay as it is.

2.1.10 Recommendations

One participant had the idea of typically used example sentences already preinstalled in the app, e.g. “Where is the next train station?”. This would decrease the amount of text users have to type. These sentences could be organised into categories like: medical, shopping, travelling, etc. Own standard sentences should be saveable as well. This idea was strongly approved by the other participants.

One participant had the idea that the video recording screen could have a coloured frame: green meaning angle, light, distance from camera okay; red meaning the input is not good, angle, distance or lightning should be changed.

Regarding the evaluation design: participants found it difficult to first play with the app and then afterwards comment on it. They would prefer to immediately give the feedback while working with the app.

The facilitator changed the task to have two separate groups of discussion and asked participants to take notes while discussing. Afterwards they presented their collected feedback to each other. This was very useful, as the notes taken by the participants could be used for the analysis of the focus group meeting.

For further recommendations see 2.1.2–2.1.9.

2.1.11 Use cases

The participants named different use cases where they would use the app:

- Medical emergencies, like going to the accidents and emergencies department at a hospital.
- Situations in public transport.
- While shopping, like asking where one can find goods in a store.

2.2 Translation

2.2.1 General feedback

The facilitator started the session with some general information about MT and the used data. He then asked if everyone is familiar with glosses, which was the case, and further explained how the example sentences are presented: First the input in gloss format, then the

MT output and then the human translation (HT). He then asked the participants to discuss the shown translation.

For the second part he explained that the MT is now in the other direction, with first shown the backtranslation of the signed utterance in the corpus, then the MT and then the original signed sentence presented in glosses.

Participants did not fully understand how the MT works. Especially the input raised a lot of questions, as it did not become clear that the corpus data is used for training, and how the training works in general. Participants were wondering how this will work in the final app – if a person can annotate their own data or not.

In general, for the direction of DGS to German the participants preferred the HT. A lot of the MT sentences made them laugh; some were just rated as being wrong. In the direction of German to DGS, participants were more pleased with the MT. In single cases they even preferred the MT to the HT as it was more specific (e.g., sentence 3, see Table 8).

TABLE 8: DE→DGS SENTENCE 3

Input	MT	Original
Das Baby war gehörlos. (The baby was deaf.)	BABY1 TAUB-GEHÖRLOS1	\$INDEX1 TAUB- GEHÖRLOS1 \$GEST-OFF

The participants stressed that the automatic translation is a hard task for scientists. They are aware of the difficulties and see the presented examples as a progress. They are also aware of shortcomings that need to be addressed (non-manual markers, prosody, etc.) but see that as further steps that will be made in the future. In general, they seemed positively surprised by how good MT already is. Comparisons with Google translate were made to discuss the progress.

In the second part of the task (DE→DGS) participants were discussing less and being distracted more easily. Some topics from task I were mentioned again or new ideas regarding the app and the use cases were added. It is hard to tell if this is due to the design (already 4 hours into the meeting, shortly before the next break) or if it could mean that the translation from German to DGS is less controversial.

2.2.2 Context

Context was the most-often mentioned topic in task II. Participants discussed the missing context which made it difficult to understand the translation. In Sentence 4 (see Table 9) for example the pronoun is missing in the glosses but known to the recipient due to context.

TABLE 9: DGS→DE SENTENCE 3

Input	MT	HT
NUM-HUNDERTER1 METER1 SPRINGEN2 STARK3	Der 100-Meter-Lauf war stark. (The 100-metre race was strong.)	100-Meter-Lauf und Weitsprung, darin ist er richtig stark. (He is really strong in the 100-metre race and the long jump.)

Other sentences would be translated completely differently by the signers as they had no context. Sentence 8 (see Table 10) would be translated by the group as “This is a nice bike”.

TABLE 10: DGS→DE SENTENCE 8

Input	MT	HT
\$INDEX1 SCHÖN1 FAHRRAD1	Das ist schön, mit dem Fahrrad zu fahren. (That's nice to ride a bike.)	Ist aber schön zum Fahrrad fahren. (But it's nice for cycling.)

It was also discussed that sometimes it is not clear what the signer means. In case of sentence 12 (see Table 11) both translations were okay for the participants, as both meanings are possible.

TABLE 11: DGS→DE SENTENCE 12

Input	MT	HT
DU1 SCHWIMMEN1 HIN- UND-HER1 ANSTRENGEND5 SEHR7	Wenn du zum Schwimmen bist, ist das anstrengend. (When you're out for a swim, it's exhausting.)	Wenn du hin- und herschwimmst, das ist doch unheimlich anstrengend. (When you swim back and forth, it's incredibly exhausting.)

2.2.3 Tense

Participants wondered where information on tense comes from as the MT and the HT use different ones in several sentences (see for example Table 12).

TABLE 12: DGS→DE SENTENCE 1

Input	MT	HT
JUNGE6 ICH2	Ich war noch jung. (I was still young.)	Ich bin jung. (I am young.)

2.2.4 Glossing

Some of the glosses were judged to be a little bit off in regard of their name/associated word. At this point it was mentioned that the qualifiers within the DGS Corpus were not used, but only the base form of the gloss. This way information is lost. The example discussed can be seen in Table 13.

TABLE 13: DGS→DE SENTENCE 2

Input	MT	HT
-------	----	----

NETT1 EINDRUCK1 STRAHLUNG2	Er war nett und hatte eine Strahlung. (He was nice and had radiation.)	Der macht schon einen ganz netten Eindruck. (He makes quite a nice impression.)
-------------------------------	--	---

The gloss \$INDEX1 (a pointing sign) was discussed as it can mean different things depending on context and where it is directed to. This information is missing completely in the glossing. One example for such a case is Sentence 2 (see Table 14).

TABLE 14: DE→DGS SENTENCE 2

Input	MT	Original
Kennst du das? (Do you know that?)	KENNEN1 DU1	KENNEN1 \$INDEX1

Other aspects that were mentioned as being missing in the glosses are: directionality of signs, mouthings, body movement.

2.2.5 Emphasis

Participants discussed that there are different possibilities to emphasize a sentence (with intonation, etc.) and that based on the written format it is not clear which part of the sentence should be stressed.

2.2.6 Missing content

Participants wondered why sometimes whole signs are not taken into the translation. For example, in Sentence 7 (see Table 15) the gesture is not added to the translation but this gesture mediates the attitude of the signer.

TABLE 15: DGS→DE SENTENCE 7

Input	MT	HT
\$GEST-ABWINKEN1 NÄCHSTE1 TAG1 \$INDEX1 ZEITUNG1 PRESSE1	Am nächsten Tag habe ich in der Zeitung gelesen. (The next day I read in the newspaper.)	Am nächsten Tag haben die Zeitungen alle davon berichtet. (The next day, the newspapers all reported it.)

Another topic that was raised is the meta discussion of signs, as seen in sentence 5 (see Table 16). The information about which sign exactly is meant is missing and participants criticised that the machine has to guess which sign for “Opa” is used in the translation.

TABLE 16: DE→DGS SENTENCE 5

Input	MT	Original
Du gebärdest so für OPA? (You sign like this for OPA?)	DU1 OPA6 DU1	OPA4 DU1

2.2.7 Non-manual markers

Especially in the translation direction from German to DGS, non-manual markers were mentioned by the participants to be extremely important as they transfer a lot of information and make the output much better once they are there. For example, Sentence 2 (see Table 14) is understood as a question only with the correct non-manual markers.

2.2.8 Sign-supported speech

In one case (sentence 9, see Table 17) the MT was evaluated as very good and close to the HT. One signer mentioned that the input for this sentence is not typical DGS but sign-supported speech. They discussed how sign-supported speech may help the MT to do a better job.

TABLE 17: DGS→DE SENTENCE 9

Input	MT	HT
ARZT1 WOLLEN2 NUR2 GELD1 VERDIENEN1 \$GEST	Der Arzt will nur Geld verdienen. (The doctor just wants to make money.)	Den Ärzten geht es nur ums Geldverdienen. (The doctors are only interested in making money.)

2.2.9 Number of hands

Within the discussions in task II the question was raised if the app will be able to translate content that is signed with only one hand. As the signer will hold the device in one hand it suggests itself to only sign with one hand, although the device could be put somewhere so that both hands are free to use.

2.2.10 Recommendations

One participant suggested to use several sentences with small variants but the same human translation to train the machine.

Participants discussed the general quality of the translation and its implication in use cases. They suggested to have some information in the app telling users that the translations might not be 100% correct. They also would like their signs to be translated to written German first before the app vocalizes the translation. This way they can check via the text if the translation is the intended meaning before other people hear it. They also suggested a function where one can add improvement proposals to translated sentences in the sense of “No, this is what I meant: **new translation**”.

Regarding the task design: This task should start with more information on MT in general. Not only which data is used, but how MT works. For this the facilitator should be better prepared to be able to answer the questions regarding MT.

2.3 Avatar

2.3.1 General feedback

The facilitator started by explaining that the following task will be done via an online questionnaire and a successive group discussion. As the pilot study showed that the query of metadata is not self-explanatory, the facilitator added two explanatory points (see Figure 46):

- Language proficiency is to be understood as follows:
 - *Anfänger:in*: Common European Framework of Reference (CEFR) levels A1 and A2
 - *Mittelmäßig*: CEFR levels B1 and B2
 - *Fortgeschrittene:r*: CEFR levels C1 and C2
 - *Erfahrene:r*: L1 signer and interpreters
- University as the place where one has learned sign language can be found under “in der Schule (at school)” and then “außerschulische Sprachkurse (extracurricular language course)”

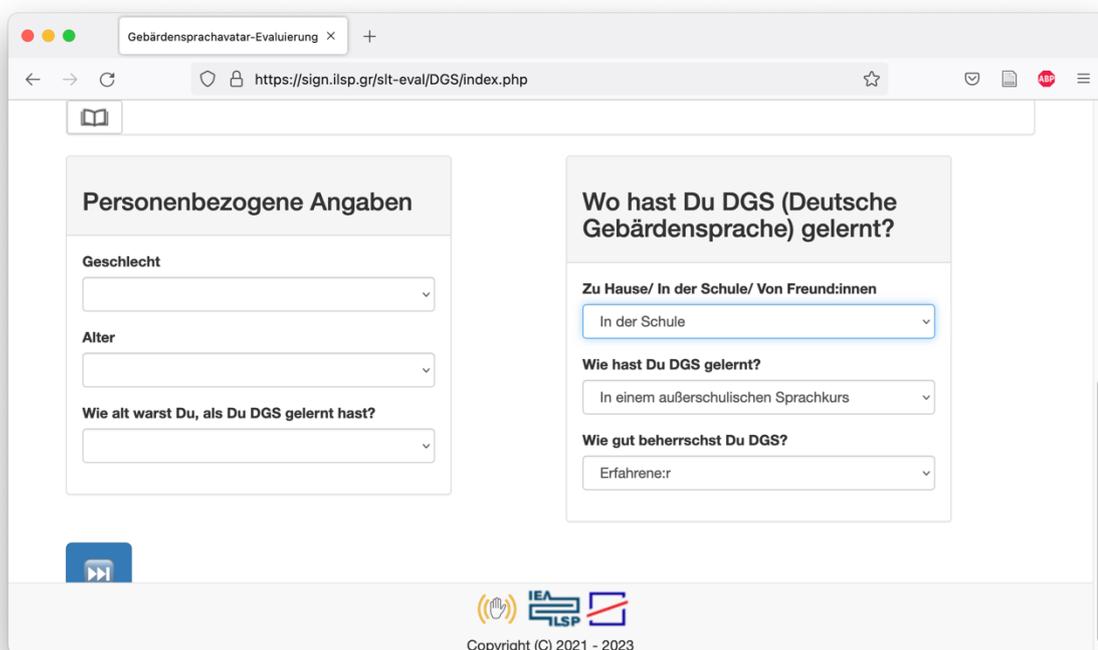


FIGURE 46: META DATA QUESTIONNAIRE

The facilitator helped with questions that raised while filling the questionnaire. The organizer (Maria) was present during the instructions and came back for the last 30 minutes of discussion.

Participants completed the questionnaire on the computers present (iMacs owned by the IDGS). They filled out the questionnaire at ~15:00 CEST.

One methodological issue that came up is that some of the participants knew the content from a feedback round that happened within the DGS corpus meetings.

In general participants stated that they saw an improvement from the avatar used in the DGS Corpus (named Anna) to Paula and they complimented this step forward. The biggest point of critique were the mouthings and mouth gestures.

2.3.2 Intelligibility

Participants described the avatar being difficult to understand. Although the signs and the meaning were the same as the ones by the human signer, they had problems understanding the avatar.

Participants assumed that they would become better in understanding the avatar when using it repeatedly in the same contexts. This was compared to meeting a new signer when it also takes time to accommodate to the style of signing. One participant said that they watched each movie in the questionnaire three times to better understand them. The first time the avatar was unintelligible, but then it became easier to read the signed content.

2.3.3 Depiction

One participant said that the avatar was too small to understand properly, e.g., reading the mouthings was hard. The participant suggested that it should be possible to zoom in on the avatar. This was also seen as a help for old people.

Another point that was raised is the angle in which the avatar is seen. It was proposed that the avatar can be moved around to have different angles, as some signs are hard to read from a front view.

The participants stressed that they would like to be able to change the speed of the avatar.

2.3.4 Mouthings and non-manuals

The participants described the avatar to be too smooth in the face. The mouthings were criticised to be too small. The participants stressed that the mouthing is important to understand the avatar and that there could be more mouthings and that they could be more intense.

Participants criticised that mouth gestures were wrong, e.g., in the case of BALD (“soon”) (see Figure 47). Participants described the sign BALD hard to understand as a whole.

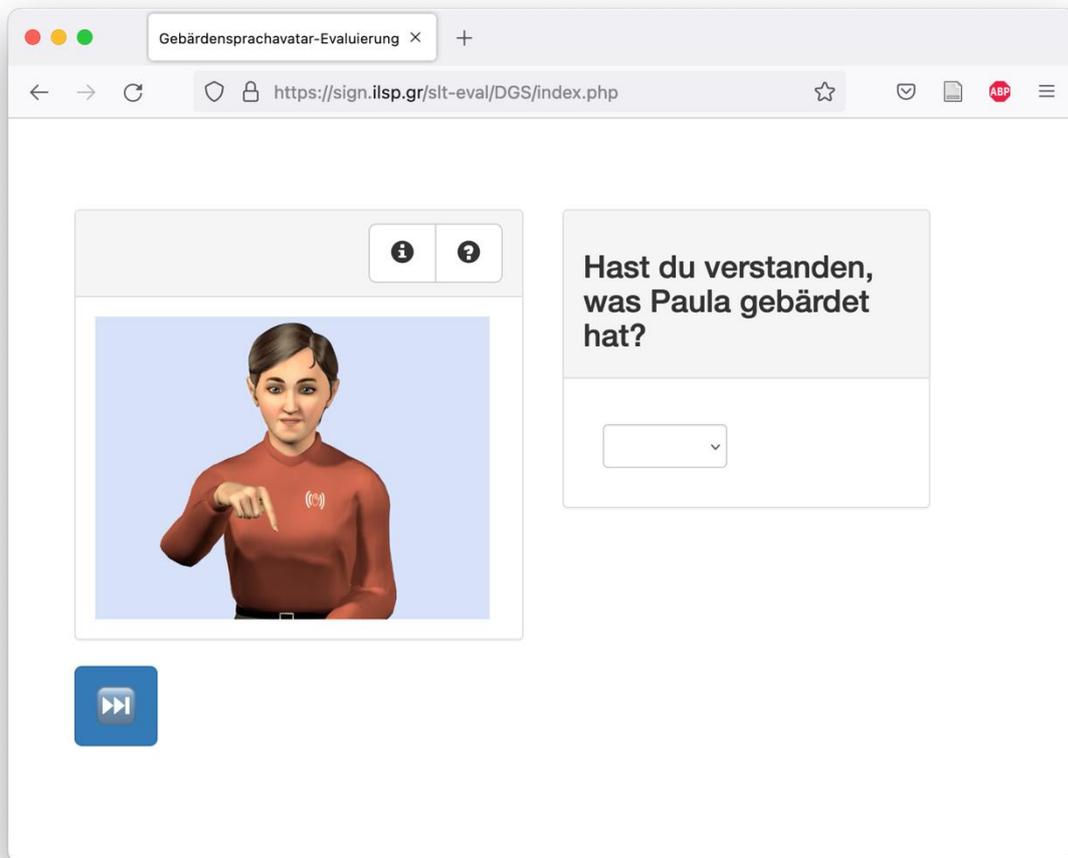


FIGURE 47: SIGN BALD WITH MOUTH GESTURE

Participants said that they could see an improvement in the mouthings compared to earlier avatars but that there is still room for improvement.

The group discussed if there should be a setting to choose how intense the mouthings should be. One participant thought that this is too much and it should be the same for everyone. Another participant liked the idea. Another participant also disliked the idea of a setting for intensity but stated that the mouthings should be more human-like.

2.3.5 Recommendations

Participants recommended to use the data stored in iLex on mouthings and mouth gestures to train the avatar (they did not know that we are already doing this).

The metadata questionnaire should be changed to make it easier to understand.

The participants would like to give more detailed feedback within the questionnaire. Instead of rating the signs from good to bad they would like to add what exactly they did or didn't like.

For further suggestions see 2.3.2–2.3.4

2.4 Final round of questions and discussion

In the last 30 minutes the organizer (Maria) joined the group and answered open questions, these were:

- Will there be an editor to manually write the input in the form of glosses?
- What will happen after the project has ended? Who will maintain the app?
- Will there be a project following EASIER or will it be prolonged?
- How can other research groups feed their data into the EASIER system?
- Can I collect my own database of signs and feed it into the EASIER pipeline?
- Is motion capture used within EASIER?
- How good is the video recognition at the moment?
- How will the broadcast data be used for MT? And is that data annotated?

The general feedback on EASIER was that it is an interesting project but that the goals seem too high for that short amount of time and that more research should be done in this area.

3 OTHER FEEDBACK

3.1 Feedback from facilitator (optional)

It was extremely helpful that some participants work in the DGS-Korpus project and are very familiar to the training data of the MT task. This was very fruitful to the discussion.

4 ANNEX

FEEDBACK PILOT STUDIES



EASIER Project | Intelligent Automatic Sign Language Translation
Grant Agreement no 101016982

EASIER Evaluation

Feedback from DGS pilot study

Collected by Maria, UHH

Two pilots conducted:

One with a hearing participant in spoken German, 19.09.2022 (hearing facilitator on site)

One with a deaf participant in DGS, 20.09.2022 (both facilitators on site)

General Feedback:

- The participant liked to be part of this evaluation.

App:

- This task was quite frustrating to our participants and they needed help navigating through the app.
- In general, there was a lot of discussion going on while testing the app, this could be due to the fact that the participant was alone with the facilitator(s).
- Special attention should be paid to explaining, that it is a click dummy and that not all technicalities are working yet. This was not clear and led to confusion.
- Works better on computer than phone (with the phone one tends to refresh the page which brings one back to the start).

MT:

- It was not fully clear what this task exactly should look like, which does not mean that this quite open discussion can't be fruitful. But maybe we should discuss once more how to introduce the task to make it less confusing for participants in the beginning.
- The deaf participant and moderator would have preferred to first only see the input and then in a second step see both translations (MT output and original) and discuss them.
- Original input sentences preferred to single gloss videos as a lot of information is missing (by deaf participant, hearing participant didn't have strong feelings in this regard)

Avatar:

- The questionnaire is not like it should be: In between the different tasks there is no explanation (or put differently the question is not asked), although we recorded

- 1 -



INVITATION MAIL

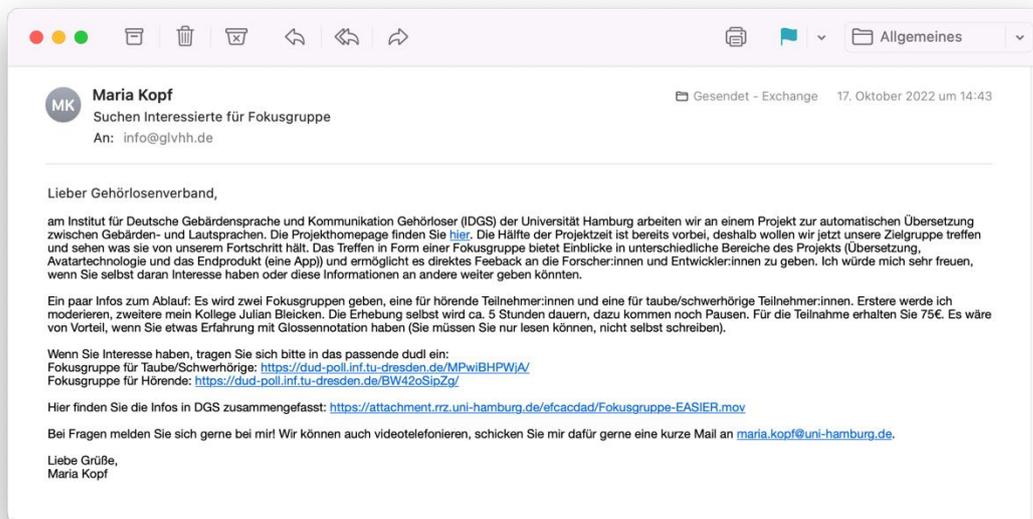


FIGURE 48: EXAMPLE FOR INVITATION MAIL

Translation into English:

Dear Deaf Association,

at the Institute for German Sign Language and Communication of the Deaf (IDGS) at the University of Hamburg we are working on a project for automatic translation between sign and spoken languages. You can find the project homepage here. Half of the project time is already over, so now we want to meet our target groups and see what they think of our progress. The focus group meeting will provide insights into different areas of the project (translation, avatar technology and the final product (an app)) and allow for direct feedback to the researchers and developers. I would be very happy if you are interested in this yourself or if you could pass this information on to others.

A few details about the process: There will be two focus groups, one for hearing participants and one for deaf/hard of hearing participants. I will moderate the former and my colleague Julian Bleicken the latter. The survey itself will last about 5 hours, plus breaks. You will receive 75€ for your participation. It would be an advantage if you have some experience with glossary notation (you only need to be able to read it, not write it yourself).

If you are interested, please sign up in the appropriate dudl:

Focus group for deaf/hard of hearing: <https://dud-poll.inf.tu-dresden.de/MPwiBHPWjA/>

Focus group for hearing people: <https://dud-poll.inf.tu-dresden.de/BW42oSipZg/>

Here you can find the information summarised in DGS: <https://attachment.rrz.uni-hamburg.de/efcacdad/Fokusgruppe-EASIER.mov>

If you have any questions, please feel free to contact me! We can also do a video call, please send me a short mail to maria.kopf@uni-hamburg.de.

Kind regards,
Maria Kopf

SLIDE SET I

Can be downloaded from: <https://attachment.rrz.uni-hamburg.de/efcacdad/Intro-Evaluation-DGS.pptx>

SLIDE SET II

Can be downloaded from: <https://attachment.rrz.uni-hamburg.de/efcacdad/Evaluation-MT-sentences.pptx>



GERMAN SIGN LANGUAGE HEARING GROUP

- **NAME OF PERSON WRITING REPORT: MARIA KOPF**
- **DATE OF EVALUATION: 28 OCTOBER 2022, 10:00–16:00 CEST**
- **SIGN LANGUAGE: GERMAN SIGN LANGUAGE**
- **GROUP (DEAF/HEARING): HEARING**

1 METHOD AND PARTICIPANTS

1.1 Pilot Study

Two pilot studies were conducted, one for the deaf and hard of hearing (HoH) group, one for the hearing group. This should ensure that both facilitators were prepared for the task. The pilot study for the hearing group was conducted on the 19.09.2022. Present was one hearing participant and the hearing facilitator/organizer (Maria). All parts of the planned study were tested. The pilot study lasted XX hours. A summary of both pilots (one for the hearing group, one for the deaf group) can be found in the Annex under *0 Feedback Pilot studies*.

Based on these findings the following adjustments were made:

- The app is tested on mobile phones or tablets instead of computers.
- Slides of the Machine Translation (MT) task will be presented stepwise to better navigate the discussion.
- The video shown at the beginning of the Avatar task is moved to the beginning of the meeting and will be skipped in the Avatar task.
- Minor changes in the slides to better support the facilitators.

1.1 Facilitator

Maria Kopf was the facilitator of the hearing group. She is a research associate at the Institute for German Sign Language and Communication of the Deaf (IDGS) at the University of Hamburg. She works for the easier project in WP6 'Resource harmonization'. Maria holds an M.A. in Sign Languages of the University of Hamburg.

Maria is hearing and learns DGS since 2018. From 2016–2018 she learned ÖGS. German is her L1, she also speaks English and Dutch.

1.2 Participants

For recruiting participants an invitation mail in written German and DGS (see Annex 0) was sent via an internal mailing list at the IDGS, a mailing list of the professional association of sign language interpreters in Northern Germany (Berufsverband der Gebärdensprachdolmetscher/-innen in Norddeutschland (BGN) e.V.), the newsletter of the Deaf Association Hamburg (GLVHH – Gehörlosenverband Hamburg) and to the interest group of Deaf students in Hamburg (iDeas Hamburg). The DGS version was not sent in the case of e-mails sent exclusively to hearing addressees. Additionally, contacts to members of the signing community in Hamburg were used to invite people personally.

Five participants were recruited. The ages ranged from 23 to 33 years old. All of the participants are hearing and female, one of them is a CODA (child of deaf adults), her L1 are German and DGS (although she stated that it was a sign supported speech way of signing and later become more and more DGS input). The other four participants named German as their L1 and learned DGS around the age of 20 years old. All participants have a background in linguistics and of them is a trained interpreter.

1.3 Procedure

1. Procedure

Both Focusgroups (deaf/HoH and hearing) tested three components of the EASIER project: the app, the machine translation and the avatar. Both groups followed the same procedure. A set of slides was used to navigate through the sessions (see Annex 0[^]) which were organised as follows:

Welcome and Introduction

- Welcome
- Presentation of the planned time schedule

Consent

- Presentation of consent form, answering of questions regarding the consent
- Signature of the consent form by participants

Presentation

- Presentation of facilitator (name, age, language background, connection to EASIER)
- Presentation of participants (name, age and language background)

Background

- Project video: <https://www.youtube.com/watch?v=imyEUqixIOU>
- Further information on EASIER
- Information on evaluation procedure

Components Test I – App

- Explanation of task
- App tested by participants individually: <https://easier-integration.nuromedia.com/>
- After testing the app freely, the participants tried to solve the following assignments:
 - Make default settings
 - Make a translation
 - Open the translation in the history/archive
- Group discussion of the app in general and the following questions.
 - How easy was it to make a translation?
 - How intuitive is the use of the app?
 - How clear is the layout of the app (visually)?
 - Which functions are still missing?

Break

Components Test II – MT

- General information on MT
- Explanation of task
- Discussion of MT sentences

Break (10 min)

- Final discussion of MT task

Components Test III – Avatar

- Explanation of task
- Online questionnaire filled in individually: <https://sign.ilsp.gr/slt-eval/>
- Group discussion of avatar

Final round of questions and discussion

2. Duration

In total the focus group meeting lasted 6 hours including breaks.

Introduction and Background: 12 minutes

Components Test I: 100 minutes
Components Test II: 100 minutes
Components Test III: 60 minutes

1.4 Technical setup

The evaluation took place in person at the IDGS. Participants sat in a half circle for general input and group discussions (see Figure 49), for the individual tasks they used computers (iMacs) facing the wall provided by the IDGS or their own phones. A large screen (behind the facilitator) was used to show the slides and task material.

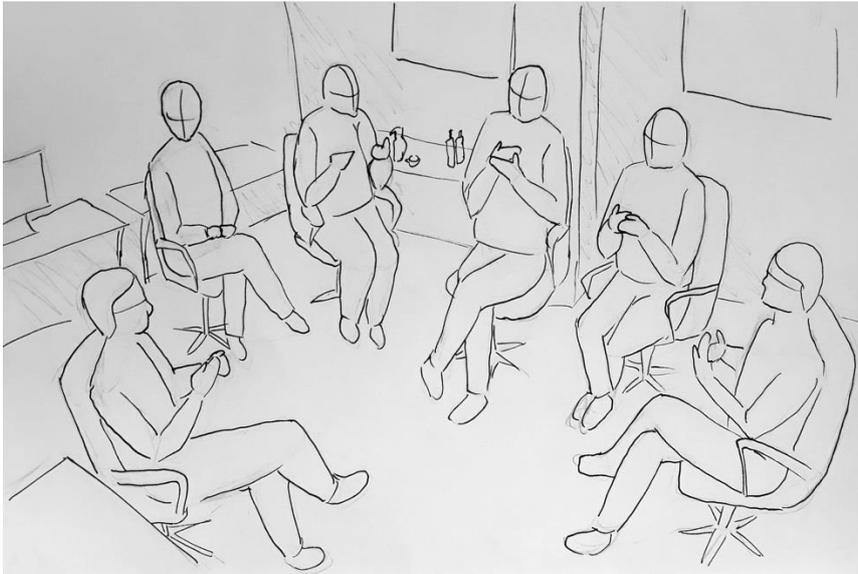


FIGURE 49: SETUP

The meeting was recorded from three different angles using three GoPro HERO6 cameras. A resolution of 1080p with 25 fps. Two cameras were set to wide angle, one to normal angle.

2 FOCUS GROUP DISCUSSION

2.1 App

2.1.1 General feedback

The facilitator first asked the participants to play with the app and try to make a translation. Later she offered them specific tasks and questions to be discussed within the group.

At first participants did not like the app. They seemed very frustrated at the first trials. Once getting more used to the navigation they were able to quickly make translations and liked it. But navigation was very frustrating to them and not intuitive at all.

Not all participants were able to make a full translation for both directions and all languages. For some participants video and audio recordings did not work.

The participants used their own smartphones with Android and iOS operating systems and Firefox, Chrome or Samsung web browsers.

2.1.2 Layout

The participants criticised that the screen is not adapting to their phone size and not all buttons are visible on their screen immediately, but that they had to scroll down or sideways to push for example the “start” button.

The drop-down menu is open when the landing page is opened (see Figure 50). This was very irritating to the participants. They also reported the bug that when the drop down is closed and opened again, it shows a subcategory of the listed points. The participants would like the drop-down menu to be in a different colour then the background.

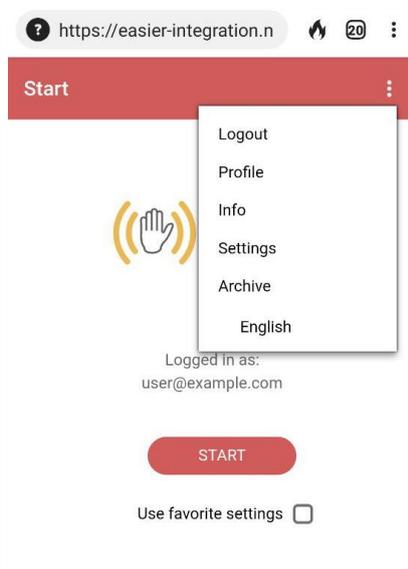


FIGURE 50: LANDING PAGE WITH DROP DOWN-MENU

Participants liked that there are options to change the layout for different font sizes, different contrasts, different colours for the avatar and the option of a dark mode. One participant even preferred dark mode to light mode, but she criticised that parts of the logo are missing in dark mode (see Figure 51).

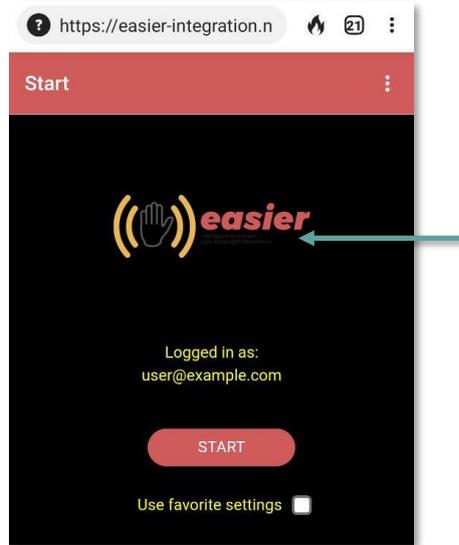


FIGURE 51: LANDING PAGE IN DARK MODE

The participants noticed that some of the buttons are not big enough to have the word in one line but with a break inside the word or overlapping each other (see Figure 52 and Figure 53).



FIGURE 52: LANGUAGE SELECTION FOR AVATAR OUTPUT

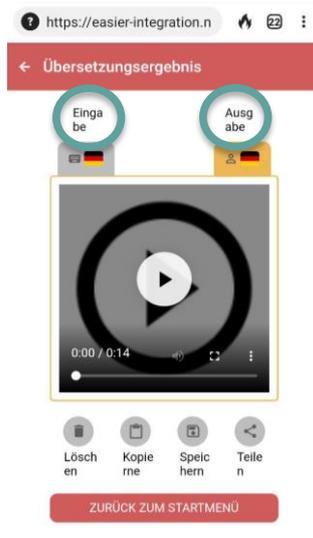


FIGURE 53: OUTPUT SCREEN

In general, they liked the colours and the simple style of the app. It was suggested though, to move the easier logo to the bottom part of the page instead of having it above. One participant especially liked the symbols for translation from question mark to phone and vice versa. Other participants were confused by these symbols and thought they have to turn their phone around or shake it.

2.1.3 Input

Participants were confused by the selection of the input and output languages. When choosing “Avatar” as output method the languages are titled with “Spoken Language” and the spoken language names are listed (see Figure 54). The same problem comes up when choosing “Camera” as input method. In this case participants thought it is expected that they would speak into the camera and not sign. They were confused by not finding an input method for signed languages. Participants suggested to only have “Language” as title and not differentiate between different modalities in the title of the page.

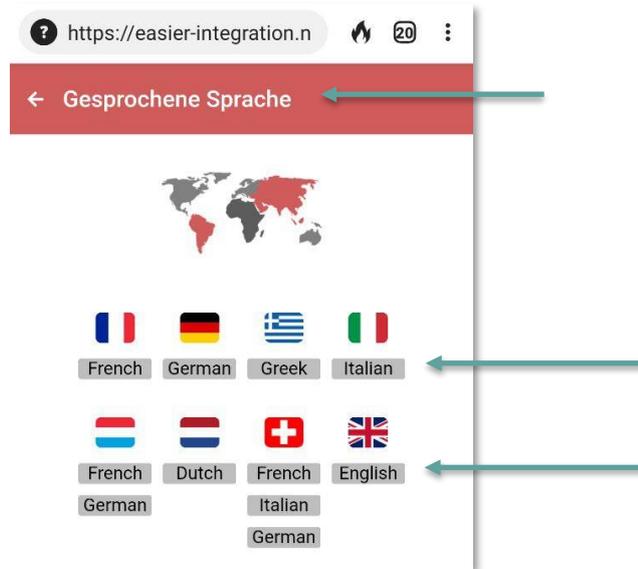


FIGURE 54: LANGUAGE OPTIONS FOR AVATAR OUTPUT

Note: To me it is not clear if one could speak into the camera or if only signing works. In the first case both – spoken and signed languages – should be listed.

The idea of commonly used sentences already prepared in the app was mentioned by the facilitator and liked by the participants. Participants suggested sentences typical for the beginning and end of a conversation. One participant suggested that the sentences are organised in categories. One of these prepared texts should be an explanation of the app itself for interlocutors that are not familiar to the app.

Participants wondered about very long input scenarios and if it would be possible to record and translate at the same time, so that for example a long talk can be recorded in several sections.

One participant had the idea to have the different inputs visible as a history comparable to a chat inside the input window.

Participants wondered if languages can be mixed in the input, e. g. Denglisch as written input or loan words in the input as well as gender fair language with asterisks or colons and such. Another question was what happens if while speaking single signs are used.

Audio input

The audio recording did not work for all participants. The question arose if there will be some visualisation of the voice being recorded. The participants would like to see some kind of visual feedback that the recording is happening. At least one participant could not see if the app is recording or not, but her phone told her that the microphone is in use. This also led to the question if microphone and camera are active permanently while using the app. Participants raised the concern that this would mean high power consumption.

Video input

For video input the participants also requested feedback in a form that one can see that the recording started and maybe also how long the recording can be. They imagined something like a bar showing that one minute can be recorded and time is running.

It was suggested that the button to start the video recording should be red and maybe beneath the video window. They would also like to be able to pause the recording when

being interrupted and play the recorded video to see if it is as intended. Participants suggested to use the commonly known video buttons: “Record”, “Pause” and “Play”. Participants wondered if one-handed signs are working with the MT and how much flexibility they have for signed input (using different objects/body parts as substitutes for the hand holding the phone). Some participants would like to turn their phone horizontally to record videos. They also wondered if they can add to the already existing video recording by overwriting only a parts of it.

Participants liked that the video input screen can be changed to full screen, but they did not like that the camera angle was not widened by this option. The button to exit the full screen mode was hidden by the cover of the phone of one participant. She suggested to move this button up a little bit.

The input video is not mirrored which was described as extremely irritating by the participants. They suggested that this could be an additional option in the setting: having the video mirrored or not.

Written input

Participants discussed the quality of the written input. They asked if the app will be able to translate the input in cases where there are spelling mistakes or wrong use of capital and small initial letters. They suggested to have words written wrong underlined, so that they can see mistakes faster. Another suggestion was to have suggestions for synonyms and a list of possible words if the word is spelled wrong, e. g. when the written word is unknown to the MT.

The maximum number of characters that can be put into the text input is not working. One participant could enter 2100 of 2000 possible characters.

One participant had the idea of a fourth input method: via the picture of a text. She compared this function to google translate and mentioned that this would be extremely helpful to deaf people. The group strongly agreed with this idea. As a use case the menu in a restaurant was named as it is an extremely quick way to translate written text.

2.1.4 Output

One participant noticed that the avatar is signing outside the screen so that the finger is cut.

The participants suggested that the output language can be changed directly in the output screen via clicking on the small flag (see Figure 55). This way one step in the navigation could be saved.

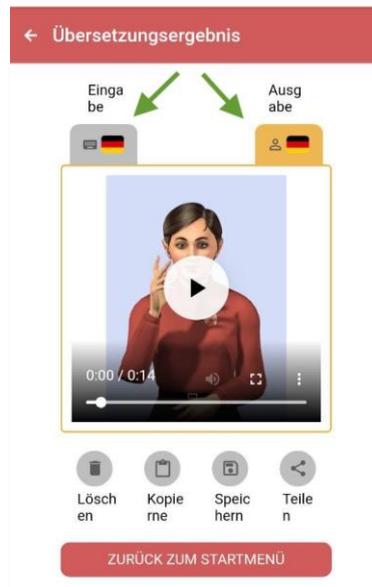


FIGURE 55: AVATAR OUTPUT

2.1.5 Dark mode

One participant noticed that the text is not readable when entering the dark mode. Another participant found the setting for font colour within the avatar settings. The place to find them seems to be not very intuitive.

When in dark mode the drop down menu from the landing page does not show the “languages” section anymore (see Figure 56) and, as mentioned, parts of the logo are missing (see Figure 51). The black part of the German flag is not distinct from the background which looked strange to the participants.

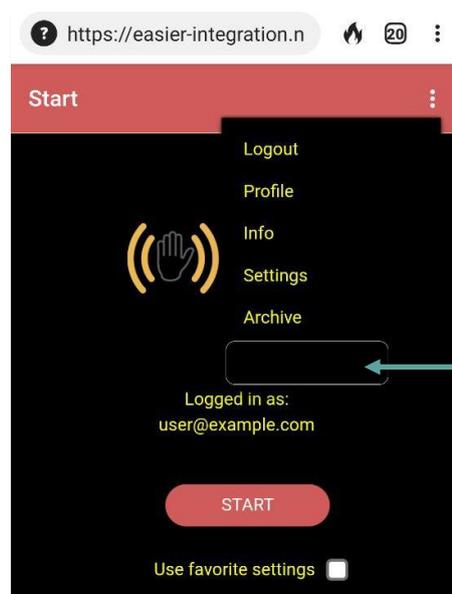


FIGURE 56: LANDING PAGE IN DARK MODE WITH DROP-DOWN MENU

2.1.6 Settings

Participants found the setting not to be self-explaining. They discussed what “Voice style” should be and what the icons (T-shirt and tie) should represent. They were also confused by the option “Adapt intonation” and wondered what effect that setting will have. They would like to have examples for gendered voice settings, like small listening examples to hear how the voice would sound.

Another participant noticed that the icon for style options with the tie is not gender fair but gendered towards male representation.

The icon for the non-binary avatar was discussed. The “X” was not liked by the participants and the facilitator added that this was already discussed within the team. Another participant emphasised that the “X” is read by her as clearly representing agender people and that the official non-binary symbol is more inclusively representing everything that is not male or female.

The participants liked that there are different setting options for the avatar (background, gender and contrast). One participant suggested to exchange with deafblind people on this topic.

One participant mentioned that it would be nice to have the option to use gender fair language (or different styles of gender fair language) for the output.

2.1.7 Navigation

Making the first translation was difficult for most participants. All of them had trouble navigating through the app. They would like selections being saved by clicking on the button and immediately being put forward, or via a “Continue” button. The need to press “Back” to go along the settings was strongly rejected. But they also mentioned that once all settings had been made the translation is done very quickly which they liked.

When using the “Backwards” button of the phone internal operating system the app is closed (as it is only a website and the back button of the phone puts one to the last opened website before the used one). Participants found this very frustrating and discussed that additionally to the “Backwards” button within the app they would like their phone internal “Backwards” button to also work.

Participants did not like that they had to click “Use favorite settings” to have their settings being saved. They would like that to automatically happen.

For selecting languages, participants did not like that they had to click on the name of the language but could not click on the flag.

Participants suggested a home-button which brings one back to the landing page.

After finishing the first translation, there should be a way to go back to the text and add a new text, instead of being thrown back to the start of the whole process and having to begin from anew. Also, the text should be saved in the input window.

As mentioned in 2.1.4 participants described the app to be very linear, in the sense that they always have to go back to the start for a new translation and are not able to change the input itself (e.g. in case of spelling mistakes), as well as input and output language and modality immediately at the output screen. They suggested further tabs within the output screen to change modalities (see Figure 55).

2.1.8 Language of the app

For several participants the app was presented in English and they had to change the settings in the first step. Even when changing the language to German not everything is presented in German: the language selection page is still in English (see Figure 54).

One participant criticised that the info text is not written in gender fair language.

One participant did not like the name of the “Start” button as it was read by her like a tour through the app and not a translation. She would prefer the word “Translate”.

2.1.9 Profile

One participant wondered how registration for a profile will work. If one has to share a lot of data and if it will be complicated. She also wondered if different profiles can connect with each other and share their translations. The participant stressed that registration should be easily accessible.

Participants also wondered where their data is being stored as it could be sensible content. It was discussed that if data is saved in a cloud solution this could be a problem with data protection but that locally saved data uses up a lot of space on the device. The participants would like this information to be transparent.

2.1.10 Archive

Participants liked that they can filter the archive for input modality.

2.1.11 Offline version

Participants wondered if there will be an offline app that one can download beforehand (e.g. when having access to Wi-Fi). One participant compared this to a dictionary app she uses which offers the service to download single languages.

2.1.12 Typos

Participants found typos in the app: “Kopierne” should be “Kopieren”.

2.1.13 Recommendations

Some of the recommendations by the participants are listed below. For further recommendations see sections 110 to 117.

- Examples within the “Voice settings” to make options clearer
- Home button
- Offline version
- More exchange with deafblind target group
- Underline spelling mistakes in written input
- Input via pictures of written text
- Change between input and output languages, modalities and content within the output screen
- Set of commonly used sentences, one of them being an explanation of the app itself

Regarding the evaluation design: participants found it difficult to first play with the app and then afterwards comment on it. They would prefer to immediately give the feedback while working with the app.

2.1.14 Use cases

The participants named different use cases where they would use the app or expect deaf people to use it (although they said that it is hard for them to tell where deaf people would use it and that they did not feel comfortable with making assumptions):

- The main use case that was named was to look up single signs (here it was mentioned that it would be nice to have specialised terms in the vocabulary of the app)
- International academic conferences
- Travelling to countries where the written language is not known
 - For visiting the restaurant
 - For doing sightseeing and reading info plates
- In medical settings like:
 - In the hospital while waiting for an interpreter
 - When staying in the hospital for several days for small interactions, like discussing meal options
 - In emergencies to communicate with ambulance staff or other present persons
 - In the first contact or registration with the accident and emergency department while waiting for an interpreter
- In public transport, especially when there are problems (here it was mentioned that it would be nice if the app could record the announcements by the train staff or the train staff would use the app to show the announcements in the train)
- When playing sports with hearing people to discuss rules

2.2 Translation

2.2.1 General feedback

The facilitator started the session with some general information about MT and the data used and answered questions on MT. She then explained how the example sentences are presented: First the input in gloss format, then the MT output and then the human translation (HT). She then asked the participants to discuss the shown translation.

For the second part she explained that MT now operates in the other direction, with the backtranslation of the signed utterance in the corpus shown first, then the MT output and then the original signed sentence presented in glosses.

After 10 minutes break the group discussed some final remarks. Within this discussion the facilitator showed the last two sentence pairs from each translation direction as these are the most complex ones within the task.

In general, the participants were surprised that most sentences made sense, for some translations they even preferred the MT output to the HT output, for example the MT output in sentence 4 (see Table 18) was seen as being closer to natural German than the HT output, which was judged to be closer to the source language.

TABLE 18: DGS→DE SENTENCE 4

Input	MT	HT
-------	----	----

DASSELBE2 FORMAT1 ERLEBNIS1 ICH1 DU1 DASSELBE2	Dasselbe habe ich auch erlebt. (I have experienced the same thing.)	Ich habe das auf die gleiche Weise erlebt. (I have experienced this in the same way.)
--	---	--

As participants discussed very fine-grained differences in meaning I assume that they were quite satisfied with the translations in general. They especially praised the quality of the translations from German to DGS. Participants emphasized that the MT output in DGS seems quite natural by also adding gestures like \$GEST-OFF.

Some questions came up during this session: The participants asked if the speed of the input will be reflected by the speed of the output (e. g. if the input is spoken very fast will the avatar sign faster?) Several questions about sign recognition were asked which could not be answered by the facilitator. And they wondered how classifier signs will be handled within MT.

2.2.2 Context

Participants mentioned that the missing context makes it harder to translate the sentences. For example, with sentence 3 (see Table 9) participants were wondering why the MT does not translate the glosses into jumping 100 meters but running 100 meters.

TABLE 19: DGS→DE SENTENCE 3

Input	MT	HT
NUM-HUNDERTER1 METER1 SPRINGEN2 STARK3	Der 100-Meter-Lauf war stark. (The 100-metre race was strong.)	100-Meter-Lauf und Weitsprung, darin ist er richtig stark. (He is really strong in the 100-metre race and the long jump.)

2.2.3 Tense

Participants wondered where information on tense comes from as the MT and the HT use different ones in several sentences (see for example Table 12). They discussed that time is hard to translate as it is often expressed in a spatial manner or not expressed at all but only through context.

TABLE 20: DGS→DE SENTENCE 1

Input	MT	HT
JUNGE6 ICH2	Ich war noch jung. (I was still young.)	Ich bin jung. (I am young.)

In the other direction (DE→DGS) it was questioned if tense should be marked if it is in the input (see for example Table 8) and if so, how that could be done. Maybe this could be

solved with the sign GEWESEN¹¹ which is described by the participants as a regional variation common to the area of Hamburg.

TABLE 21: DE→DGS SENTENCE 3

Input	MT	Original
Das Baby war gehörlos. (The baby was deaf.)	BABY1 TAUB-GEHÖRLOS1	\$INDEX1 TAUB- GEHÖRLOS1 \$GEST-OFF

One participant mentioned that she would expect that within the use case scenarios present tense will more likely be used than past tense as it is a situation of direct exchange.

2.2.4 Intended meaning

Participants wondered if the translation always was in line with the intended meaning. In case of sentence 1 (see Table 12) the particle “noch” is adding more content than can be seen by the glosses. One participant called it a “nostalgic” mood, another one read it as an evaluation of the age.

In the case of pronouns participants wondered how the MT decides which pronoun to use when translating from DGS to German as pronouns in DGS are not gendered, while they are in German (see for example Table 13).

TABLE 22: DGS→DE SENTENCE 2

Input	MT	HT
NETT1 EINDRUCK1 STRAHLUNG2	Er war nett und hatte eine Strahlung. (He was nice and had radiation.)	Der macht schon einen ganz netten Eindruck. (He makes quite a nice impression.)

Participants wondered if expressions will be translated according to the intended meaning or literally, e. g. will the English input “it is raining cats and dog” be translated into DGS: *RAIN CAT DOG* or *RAIN HEAVILY*.

2.2.5 Glossing

Based on the glosses it was sometimes hard for the participants to judge if the translation is the right one, e. g. in sentence 1 (see Table 12) based on the glosses the participants would translate the sentence as “Ich bin ein Junge. (I am a boy).”.

Based on the gloss EINDRUCK1 several aspects of glosses were discussed one of them being the missing directionality in the gloss representation of signs. One participant wondered if the agreement auxiliary PAM (personal agreement marker; glossed as AUF-PERSON1 in the DGS Corpus¹²) could be used to compensate for missing spatial marking of

¹¹ See https://www.sign-lang.uni-hamburg.de/meinedgs/types/type17167_de.html#type91079

¹² See https://www.sign-lang.uni-hamburg.de/meinedgs/types/type15599_de.html

verbs. Another participant was wondering if reference tracking is used within MT to solve the problem of directionality.

Another shortcoming of the glosses discussed are number signs. As the DGS Corpus uses qualifiers to specify numbers, e. g. “fourteen” being glossed as \$NUM-TEEN1:4 but these glosses are not used for training (used is only the shortened form \$NUM-TEEN1) it is impossible to translate numbers correctly (see Table 23).

TABLE 23: DGS→DE SENTENCE 17

Input	MT	HT
ICH1 HIER1 SCHULE1 \$NUM-TEEN1 \$NUM- ZEHNER1 \$NUM-ORD1 SCHULE1 \$GEST-OFF	Ich war in der Schule 1989 in der Schule. (I went to school in 1989 in school.)	Ich bin also 1950 in Köln in die Schule gegangen, das war meine erste Schule. (So I went to school in Cologne in 1950, that was my first school.)

Further aspects that were mentioned as being missing in the glosses are: body shifts, non-manual markers and intonation.

The gloss \$INDEX1 (a pointing sign) was discussed as it can mean different things depending on context and where it is directed to. Compared to the deaf/HoH focus group the hearing participants did not criticise this in the following example (Table 14), but thought that it is a good translation:

TABLE 24: DE→DGS SENTENCE 2

Input	MT	Original
Kennst du das? (Do you know that?)	KENNEN1 DU1	KENNEN1 \$INDEX1

The participants added that they would guess that the \$INDEX1 in the Original is in fact a DU1. They shared their experience with annotating in the DGS Corpus as student workers and reported that the gloss \$INDEX1 is much more used than DU1.

Sentence 6 (see Table 25) on the other side was criticised as the meaning depends on the direction of the index, an information missing in the glosses.

TABLE 25: DE→DGS SENTENCE 6

Input	MT	Original
Ich wohne da in der Nähe. (I live near there.)	ICH1 WOHNUNG1 NAHE1 \$INDEX1	ICH1 WOHNUNG1 NAHE2

Although one participant mentioned that this ambiguity is the same in the input sentence and therefore the translation seems suitable.

2.2.6 Missing content

Participants wondered why sometimes whole signs are not taken into the translation or why some words appear although not being in the glosses, e. g. “read” in sentence 7 (see Table 15). In the same sentence the gesture is not added to the translation although this gesture mediates the attitude of the signer, which was criticised by the participants.

TABLE 26: DGS→DE SENTENCE 7

Input	MT	HT
\$GEST-ABWINKEN1 NÄCHSTE1 TAG1 \$INDEX1 ZEITUNG1 PRESSE1	Am nächsten Tag habe ich in der Zeitung gelesen. (The next day I read in the newspaper.)	Am nächsten Tag haben die Zeitungen alle davon berichtet. (The next day, the newspapers all reported it.)

Another topic that was raised is the meta linguistic discussion of signs, as seen in sentence 5 (see Table 16). The information about which sign exactly is meant is missing and participants criticised that the machine has to guess which sign for “Opa” is used in the translation.

TABLE 27: DE→DGS SENTENCE 5

Input	MT	Original
Du gebärdest so für OPA? (You sign like this for OPA?)	DU1 OPA6 DU1	OPA4 DU1

Another thing missing in the translation is negation (see Table 28). Participants discussed if negation should happen manually or non-manually. One participant suggested that the reason for the missing negation in the MT output is due to non-manual negation in the training data which is not glossed but translated. In this respect the manual signs seem to be okay, but the problem is that the non-manual negation gets lost as the avatar does not get the information if it is not glossed and therefore will produce a sentence lacking the negation.

TABLE 28: DE→DGS SENTENCE 7

Input	MT	Original
Kein einziger Gehörloser hat dort gearbeitet. (Not a single deaf person worked there.)	\$NUM-EINER1 EINZIG1 TAUB-GEHÖRLOS1 ARBEITEN1 \$INDEX1	KEIN3 TAUB-GEHÖRLOS1 BEREICH1

Participants discussed that if the MT output chooses body anchored signs where the original was signed with a sign inflected for person the information on the person will be lost if there is no pronoun added. In the case of sentence 8 (see Table 29) the MT could also mean “I have something important to tell.”.

TABLE 29: DE→DGS SENTENCE 8

Input	MT	Original
-------	----	----------

Ich muss dir etwas Wichtiges sagen. (I have to tell you something important.)	WICHTIG1 MUSS1 SAGEN1 WICHTIG1	ICH1 BESCHEID1 WICHTIG1 WAS1
---	-----------------------------------	---------------------------------

2.2.7 Wording

Based on the gloss TAUB-GEHÖRLOS1 participants discussed the translation of signs that are under meta-linguistic discussion within the signing community in respect to political correctness. They suggested to offer several translational equivalents (“taub” and “gehörlos”) that one can choose or to even have a general setting to translate specific signs with always the same term. In the other direction (DE→DGS) it was mentioned that it would be nice if one can choose between different signs, in the case that one or more signs are seen as politically incorrect by the user. The participants suggested that there should exist a list of signs where users can decide which sign will be used. Although this would raise the problem for fixed saying using one of the signs, that is set by the user to not be shown, like SAY in the case of “hard to say”. If the user chooses the option to translate the German word for “say” always with SIGN the saying “hard to say” would not be understood anymore. In the further discussion it was mentioned that if the most frequent signs are the default signs used by the MT there will be some cases where signs seen as politically incorrect by some signers will appear, e. g. in the case of the sign for “woman”.

The participants discussed if regional variation will be taken into account and if it can be chosen which regiolect should be in the output.

Participants wondered if incorporation is taken into account or if this information will be split up into multiple signs as in sentence 4 (see Table 30). Another example that was named was the description of big eyes.

TABLE 30: DE→DGS SENTENCE 4

Input	MT	Original
Der Mann war sofort tot. (The man was immediately dead.)	MANN1 SOFORT2 TOD2	MANN7 TOD2

2.2.8 Recommendations

The participants suggested to have several options for one translation which the user can then choose from. Although they also mentioned that in a conversation this could take too much time and that this is not possible for use cases where the output is in a language that is not known to the user.

For further suggestions see sections 119 to 123.

On the task design: Participants described this task as exciting and were discussing the sentences in great detail. But they also mentioned that they were not sure if this is what EASIER is looking for. One participant stated that she would rate the quality of the translation differently if it was not on gloss level but videos of the avatar vs. the original signing.

Note: Both groups did not discuss all sentences as time was too short.

2.3 Avatar

2.3.1 General feedback

The facilitator started by explaining that the following task will be done via an online questionnaire and a successive group discussion. As the pilot study showed that questions on metadata are not self-explanatory, the facilitator added two explanatory points (see Figure 46):

- Language proficiency is to be understood as follows:
 - *Anfänger:in*: Common European Framework of Reference (CEFR) levels A1 and A2
 - *Mittelmäßig*: CEFR levels B1 and B2
 - *Fortgeschrittene:r*: CEFR levels C1 and C2
 - *Erfahrene:r*: L1 signer and interpreters
- University as the place where one has learnt sign language can be found under “in der Schule (at school)” and then “außerschulische Sprachkurse (extracurricular language course)”

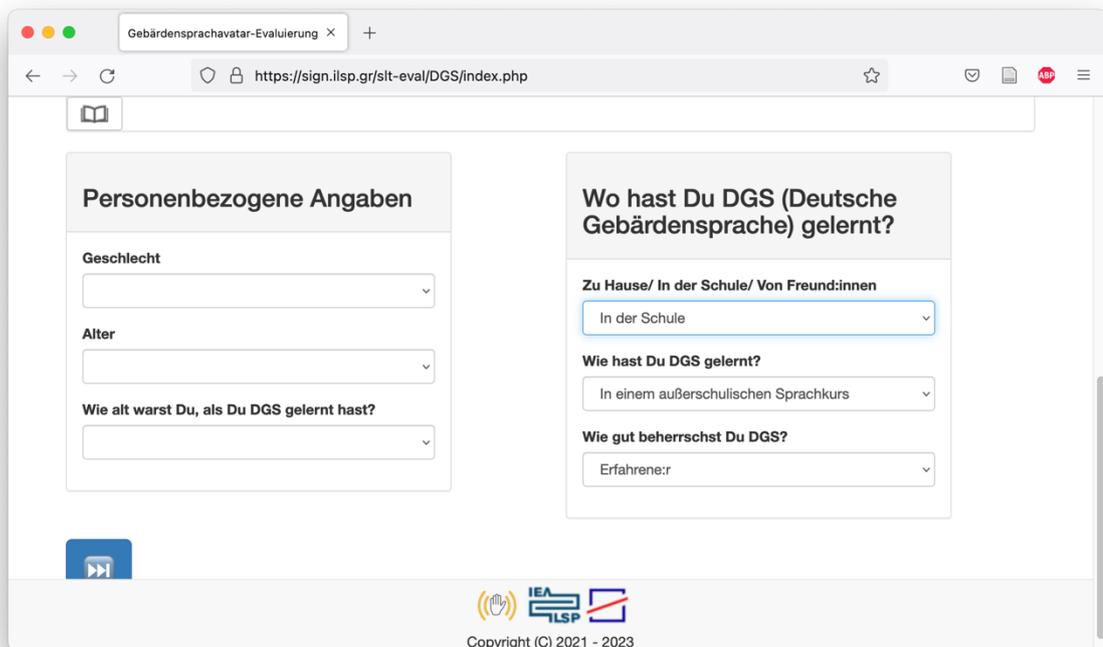


FIGURE 57: META DATA QUESTIONNAIRE

Participants completed the questionnaire on the computers present (iMacs owned by the IDGS). They filled out the questionnaire at ~15:00 CEST. The facilitator helped with questions that raised while filling the questionnaire. In the discussion afterwards the facilitator showed separate videos of Paula using the manual alphabet as the participants asked how that looks.

Participants were positively surprised how good the avatar is signing compared to older versions of avatars they know. For more detailed feedback see the following sections.

One topic that came up was the handedness of Paula and the question if this could be changed easily by just mirroring her. If it would be as easy as that the participants would like this option in the app (Paula with right hand as dominant hand and left hand as dominant hand).

Note: The glosses in the following sections align with the glosses used for preparing this task. They can be found here:

<https://docs.google.com/spreadsheets/d/13feeeYbhahDP0aThJVGP0TGpgLTkrGKejni3cHEdqTU/edit#gid=612040177>

2.3.2 Mouthing and mouth gestures

Participants stated that the biggest problem with understanding Paula is the mouthing. They discussed the case of SERVICE2A (“service”): one participant did not understand the sign at all due to the mouthing, another one got it wrong and saw “server” on the mouth and therefore interpreted the sentence wrong.

One person described the mouthings as very different from real human mouthings and specified that the timing is not appropriate. She said that they start and end too early.

Another participant said that HALLO1 (“hello”) and TSCHÜSS1 (“bye”) seemed to have the same mouthing which was irritating to her.

Participants criticised that mouth gestures were wrong, e.g. in the case of BALD1A (“soon”) (see Figure 58). In the further discussion they mentioned that if people use the EASIER app to learn new vocabulary, mistakes like that would be learned and may spread. The question arose what kind of impact that could have on DGS.

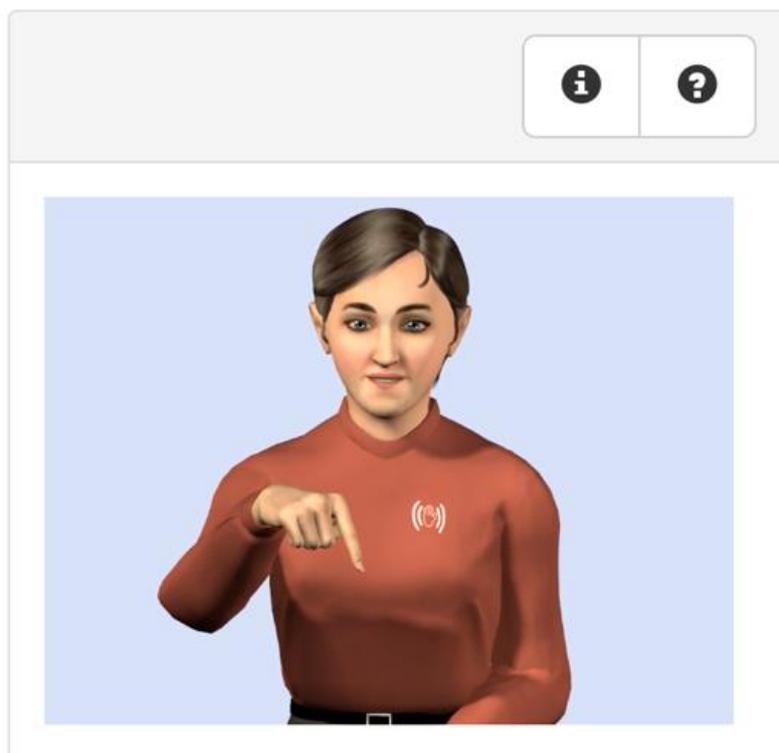


FIGURE 58: PAULA SIGNING “SOON” (BALD1A)

Participants discussed that in DGS not all mouthings are fully produced but sometimes only the stem of a word is produced. They suggested that research-based knowledge could be used to make the mouthings more human.

2.3.3 Other non-manual markers

One participant positively highlighted the movement of the eyebrows, but she found the eye aperture and departure not intense enough, especially when compared to the human signer.

Another participant highlighted positively that the body of the avatar is moving while signing which makes it look more human. The same participant also reported that the origin of the movement seemed very human, e. g. when waving the movement came from the wrist but not from the elbow or shoulder.

2.3.4 Motion

Note: the following points are partly contradictory. Paula was described as too smooth, too choppy and robotic within the same discussion, sometimes by the same participants.

Participants described Paula's motion as "too smooth", meaning that signs are fused too much in one another so that they could not define where one sign ends and the next starts. Also, it made single signs hard to read for them, e. g. the sign KANN2A ("can"). The sign EINTRETEN1A-\$SAM ("use") on the other hand was described as "too choppy".

The sign UNSER1A was described as too distal. One participant said that she did not recognise the sign when signed by the avatar.

The sign GRENZE1A-\$SAM was described to be too linear. Participants said that it lacks the movement in the wrist at the end of the sign.

In general, they described Paula's motion as a little bit robotic. To them, it looked like Paula could only use three paths as in a diagram: The x, y, and z axis without really combining them but rather moving along these sequentially, e. g. first a forward movement, then a movement to the right instead of combining these movements.

The participants also discussed intonation: The breaks between different signed utterances were criticised as being too long, e. g. between VIEL1C DANKE1 ("thank you") and TSCHÜSS1 ("bye") or in the case of BITTE1A WARTEN3 ("please wait") and BALD1A ANTWORT1 ("response is pending").

2.3.5 Appearance

One participant especially highlighted that she did like the contrast of clothing, background and skin. She was very surprised that the nails had a slightly different colour than the skin. When looking at the nails collectively the participants agreed that Paula has very beautiful nails.

The participants liked that the arms and hands throw shadows on the body although the shadow on the elbow was confusing to one participant. To her it looked like the arm is cut off and she was very irritated (see Figure 59).

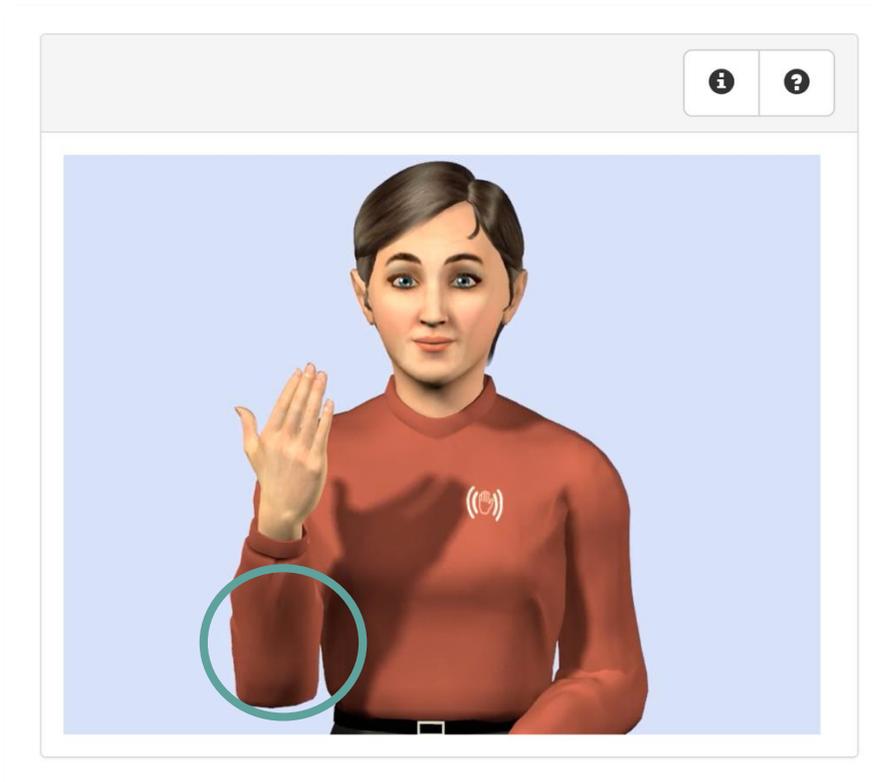


FIGURE 59: PAULA SIGNING "THANK YOU" (DANKE1)

The participants liked that the EASIER logo is on Paula's cloths and they compared Paula's appearance to Star Trek characters.

2.3.6 Manual alphabet

Participants were extremely impressed by Paula's ability to use the manual alphabet. The facilitator showed one clip where Paula spells "A-N-N-A" and another one where she spells "M-I-C-H-A-E-L" in LSF. The participants liked the trajectory of the hand while spelling and described it as very human-like.

2.3.7 Comparison to Anna

Participants compared Paula with Anna¹³ (see Figure 60). They all rated Paula's signing much better than Anna's. Especially joint movement and hand/body contact were praised to be better in Paula than in Anna.

¹³ The avatar Anna was developed in the projects *ViSiCAST*, *eSIGN* and *Dicta-Sign*.



FIGURE 60: AVATAR ANNA

2.3.8 Recommendations

Participants only recommended to have the option of choosing the dominant hand for the avatar. No other recommendations were uttered, but based on the critique described in sections 2.3.2 to 2.3.7 one can draw conclusions on the following recommendations:

- Intensify eye aperture and departure
- Shorten breaks between and within utterances
- Change length and position of mouthings
- Advance mouthings and mouth gestures (not clear how exactly)

Regarding the task design: Participants were confused that one task within the questionnaire was reduplicated. They did not see the fine-grained difference between the questions (“Does Paula sign like a human?” – “How well does she sign?” (page 4) and “Did both of them sign the same?” (page 13)). One participant added that for her the rating of “How well does she sign?” is automatically in comparison to human signing so the questions seemed all the same. She suggested to ask both questions at the same time: “How well does she sign?” and “Does she sign like a human?”. This way participants would have to separate between human-likeness and good in a more general rating. Another participant stated that if she would be completely strict in her ratings, she would have to answer all examples with a “no” as the avatar never signs in a human way.

The latter question (“Did both of them sign the same?”) was additionally confusing to a lot of participants as they did not know if it was to be understood as the same content or the exact same signs, as the signs differed in form but not in meaning.

Participants were wondering if it was on purpose that one question was asked twice (“Sorry I didn’t understand” in the part where Paula and a human signer are compared). They thought it was either a mistake or a test within the test to see if they answer consistently.

Three participants reported that they were much stricter in rating Paula after having seen the human signer signing the same sentences in comparison. Another participant reported that she was less strict after having seen the human signer as she understood the sentences much easier after that. She emphasized that there is a learning effect during the task.

3 OTHER FEEDBACK

3.1 Feedback from facilitator (optional)

In total the tasks worked better than was expected by the organizer (Maria). The participants liked the tasks and were very actively involved in the discussions. A total length of six hours (including breaks) was still okay with the participants, although both groups preferred to have one long break (one hour) and 1-2 smaller ones (10 minutes), instead of 2-3 breaks of 30 minutes.

The analysis showed that two cameras are sufficient as long as one of them captures all participants and one captures the screen with the slides and the facilitator. As backing up data takes a lot of time this should be taken into account when planning breaks. Another solution is to work with several SD-cards which can be exchanged quickly.

Some remarks on future evaluations:

- In the planning process more time should be given to invite participants and find a date for the evaluation. Especially interpreters have a fully booked schedule which makes it impossible to find a date on short notice.
- The participants that took part in this evaluation were very interested to be part in the next evaluation to see the progress. This could be taken into account when planning the following evaluation tasks.
- From the organizer's point of view communication between the different evaluating groups (different project partners) should be improved.
- The participants had a lot of technical questions. By answering them the acceptance towards new technology may be improved and knowledge within the community can be spread. This demands that the facilitators are better trained on the technical aspects of EASIER.

4 ANNEX

FEEDBACK PILOT STUDIES



EASIER Project | Intelligent Automatic Sign Language Translation
Grant Agreement no 101016982

EASIER Evaluation

Feedback from DGS pilot study

Collected by Maria, UHH

Two pilots conducted:

One with a hearing participant in spoken German, 19.09.2022 (hearing facilitator on site)

One with a deaf participant in DGS, 20.09.2022 (both facilitators on site)

General Feedback:

- The participant liked to be part of this evaluation.

App:

- This task was quite frustrating to our participants and they needed help navigating through the app.
- In general, there was a lot of discussion going on while testing the app, this could be due to the fact that the participant was alone with the facilitator(s).
- Special attention should be paid to explaining, that it is a click dummy and that not all technicalities are working yet. This was not clear and led to confusion.
- Works better on computer than phone (with the phone one tends to refresh the page which brings one back to the start).

MT:

- It was not fully clear what this task exactly should look like, which does not mean that this quite open discussion can't be fruitful. But maybe we should discuss once more how to introduce the task to make it less confusing for participants in the beginning.
- The deaf participant and moderator would have preferred to first only see the input and then in a second step see both translations (MT output and original) and discuss them.
- Original input sentences preferred to single gloss videos as a lot of information is missing (by deaf participant, hearing participant didn't have strong feelings in this regard)

Avatar:

- The questionnaire is not like it should be: In between the different tasks there is no explanation (or put differently the question is not asked), although we recorded

- 1 -



INVITATION MAIL

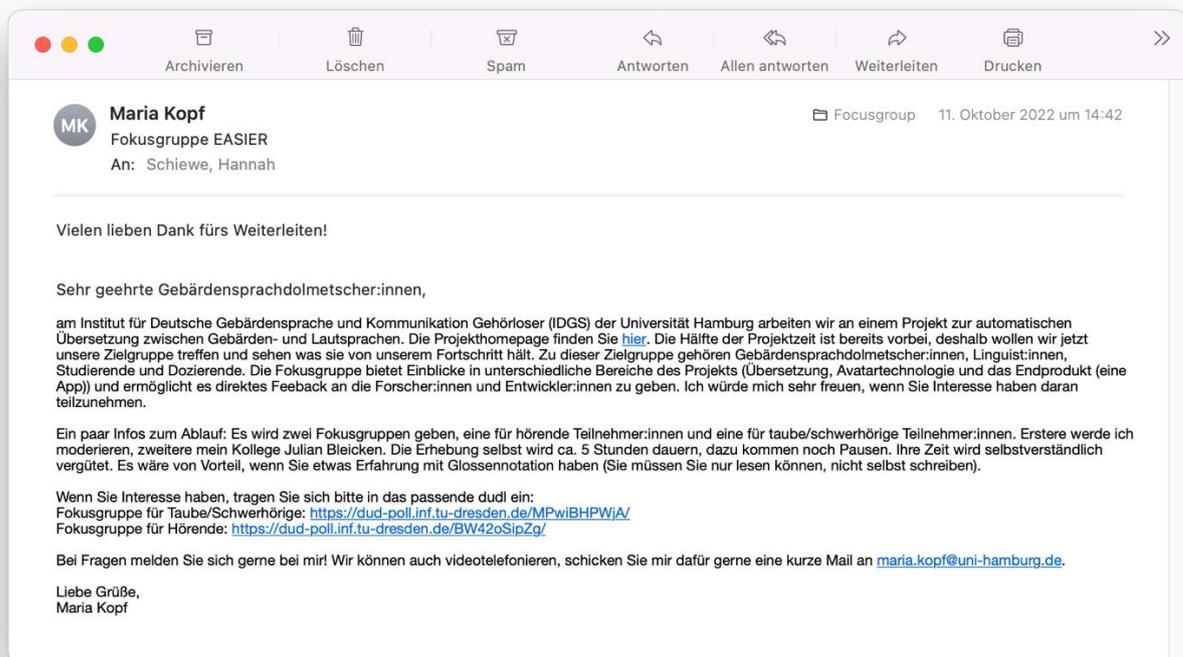


FIGURE 61: EXAMPLE FOR INVITATION MAIL

Translation to English:

Thank you very much for forwarding it!

Dear Sign Language Interpreters,

At the Institute for German Sign Language and Communication of the Deaf (IDGS) at the University of Hamburg we are working on a project for automatic translation between sign and spoken languages. You can find the project homepage here. Half of the project time is already over, so now we want to meet our target group and see what they think of our progress. This target group includes sign language interpreters, linguists, students and lecturers. The focus group will provide insights into different areas of the project (translation, avatar technology and the final product (an app)) and allow for direct feedback to the researchers and developers. I would be very happy if you are interested in participating.

Some information about the process: There will be two focus groups, one for hearing participants and one for deaf/hard of hearing participants. The first one will be moderated by me, the second one by my colleague Julian Bleicken. The survey itself will take about 5 hours, plus breaks. Your time will of course be remunerated. It would be an advantage if you have some experience with gloss notation (you only need to be able to read it, not write it yourself).

If you are interested, please sign up to the appropriate dudl:

Focus group for deaf/hard of hearing: <https://dud-poll.inf.tu-dresden.de/MPwiBHPWjA/>

Focus group for hearing people: <https://dud-poll.inf.tu-dresden.de/BW42oSipZg/>

If you have any questions, please feel free to contact me! We can also video call, feel free to send me a short mail at maria.kopf@uni-hamburg.de for that.

Kind regards,
Maria Kopf

SLIDE SET I

Can be downloaded from: <https://attachment.rrz.uni-hamburg.de/efcacdad/Intro-Evaluation-DGS.pptx>

SLIDE SET II

Can be downloaded from: <https://attachment.rrz.uni-hamburg.de/efcacdad/Evaluation-MT-sentences.pptx>



SWISS GERMAN SIGN LANGUAGE DEAF GROUP

- **NAME OF PERSON WRITING REPORT: TOBIAS HAUG & SANDRA SIDLER-MISEREZ**
- **DATE OF EVALUATION: OCTOBER 7, 2022**
- **SIGN LANGUAGE: SWISS GERMAN SIGN LANGUAGE**
- **GROUP (DEAF/HEARING): DEAF**

1 METHOD AND PARTICIPANTS

1.1 Facilitator

The evaluation was facilitated by Sandra Sidler-Miserez, who is external to the University of Zurich. Sandra Sidler-Miserez is a deaf native signer of Swiss German Sign Language (DSGS), a trained sign language teacher and has been involved in a number of research projects in the past at the University of Zurich and the University of Teacher Education in Special Needs Zurich (HfH).

1.2 Participants

Four deaf sign language teachers participated in this study. They were recruited through the professional and personal networks of faculty of the sign language interpreting program at HfH in German Switzerland. Three of the four sign language teachers were female, one was male. The age range was 40 to 59 years. No one else was present during the evaluation session.

1.3 Procedure

The evaluation was conducted in person at HfH. The participants were instructed about the EASIER project in general and about the purpose of the evaluation studies. Informed consent forms and fact sheets about the EASIER project (available in German and DSGS) had been created beforehand (DSGS version of fact sheet [here](#), of informed consent form [here](#)) and were signed prior to the evaluation study. The consent forms involved permission to use video and audio recordings of the focus group sessions as well as the responses given in the online survey on the avatar for research purposes in anonymized form (without audio or video; using pseudonyms instead of participants names). The evaluation took place on October 7, 2022 in the morning and consisted of the following parts:

- Overall introduction and instructions regarding the evaluation in DSGS and German text
- Instructions regarding evaluation of the app
- Testing the app (each participant on own laptop)
- Focus group discussion about the app with all participants
- 15-minute break
- Instructions regarding evaluation of the avatar
- Filling in avatar survey (each participant on own laptop)
- Focus group discussion about the avatar with all participants

The study session was video-recorded for later analysis.

1.4 Technical setup

The participants were instructed to bring their own laptop. Participants were seated at a different desk when working on their own. During the focus group parts of the evaluation, they were seated in a circle. The focus group was recorded with three video cameras from different angles.

2 FOCUS GROUP DISCUSSION

APP

1.1.1 General feedback

The general feedback was mixed: Overall, the navigation of the app was not considered as very user-friendly, and clarity of instructions in the app was found missing. All participants reported that they found this evaluation task challenging since the translation component was not yet implemented.

1.1.2 Navigation

The first topic that came up during the focus group was that the deaf participants did not find the navigation of the app intuitive and user-friendly. They tried to navigate through the app (e.g., language buttons, video recording), but did not, for example, find the settings, logout and profile options (top right), which would have helped to explore the full potential of the app. Additionally, navigating back (back arrow on the top left) was found to be rather confusing than conducive the experience of the app.

Participants also reported that the overall goal and use of the app was not clear to them (a possible explanation is that the missing translation module in the app hindered the evaluation of the full potential of the app).

1.1.3 Recommendation

The participants suggested a design for the translation interface similar to that of DeepL, where there are two adjacent windows for the source (left) and the target language (right). They considered the windows for typing text very small right as they are at the moment. Additionally, participants suggested that the process of selecting the source and the target languages of the translation process could be made easier by displaying logos for source and target language one below the other. Participants also suggested to use clearer logos for the different language modalities.

AVATAR

1.1.4 General feedback

The general feedback regarding the avatar was mixed. Some participants did not deem the signing of the avatar natural.

1.1.5 Linguistic issues

One of the topics that came up during the focus group interview were the non-manual features of the avatar: The participants found that the mouthing was mostly not very clear, and that general facial expression was often missing. They also deemed the signing rhythm to be not quite natural (i.e., no change of signing speed and lack of prosodic markers to structure an utterance) and observed a lack of movement of the upper torso. Eyegaze was perceived as “staring” at one point, and mention was made that eyegaze didn’t change

during signing at one point. One participant suggested that the eyegaze looked as if the avatar squinted.

The lack of prosodic markers also had an influence on the comprehension of the deaf signers, i.e., they only understood a sentence when they double-checked with the sentence of the human signer. One participant reported that when the avatar uses space (e.g., signing something on right or left side of the signing space), the movement of the upper torso (i.e., turning from the right to the left side) was choppy.

When the avatar used an index (i.e., establishing a referent), it was not clear to what or whom the index is referring to. One participant talked about a sequence where the avatar signed ANTWORTEN (to respond) from a higher position in space as he/she would have expected in discourse; to the participant it was not clear why the sign was produced higher up.

1.1.6 Technical issue with survey

One participant reported that s/he could not see her recording (only black screen), but the remaining three participants could.

2 OTHER FEEDBACK

FEEDBACK FROM FACILITATOR (OPTIONAL)

The evaluation of the app in the absence of the translation module was difficult for the participants, even though they received clear instructions (i.e., it was not clear to the participants what the goal of the evaluation was).



SWISS GERMAN SIGN LANGUAGE HEARING GROUP

- **NAME OF PERSON WRITING REPORT: TOBIAS HAUG**
- **DATE OF EVALUATION: OCTOBER 12, 2022**
- **SIGN LANGUAGE: SWISS GERMAN SIGN LANGUAGE**
- **GROUP (DEAF/HEARING): HEARING**

1 METHOD AND PARTICIPANTS

1.1 Facilitator

The evaluation was facilitated by Tobias Haug, who is external to the University of Zurich. He holds a permanent position at the University of Teacher Education in Special Needs (HfH). Tobias Haug is hearing, he has studied sign language linguistics at Hamburg University and deaf education in Boston. He holds a Ph.D. in sign languages. He is the director of the BA sign language interpreting program at HfH and learned different sign languages as an adult.

1.2 Participants

Four hearing Swiss German Sign Language (DSGS)/German interpreters participated in this study, two of whom are also instructors in the sign language interpreting program at the HfH. They were recruited through the professional networks of sign language interpreters in German Switzerland. All four were female and between 45 to 60 years old. No one else was present during the evaluation session.

1.3 Procedure

The evaluation was conducted online via Zoom. The participants were instructed about the EASIER project in general and about the purpose of the evaluation studies. Informed consent forms and fact sheets about the EASIER project (available in German and DSGS) had been created beforehand (DSGS version of fact sheet [here](#), of informed consent form [here](#)). The consent forms involved permission to use video and audio recordings of the focus group sessions as well as the responses given in the online survey on the avatar for research purposes in anonymized form (without audio or video; using pseudonyms instead of participants names). They were signed by the study participants prior to the study session. The evaluation took place on October 12, 2022 in the morning and consisted of the following parts:

- Overall introduction and instructions regarding the evaluation
- Instructions regarding evaluation of the app
- Testing the app (each participant for herself)
- Focus group discussion about the app with all participants
- 15-minute break
- Instructions regarding evaluation of the avatar
- Filling in avatar survey (each participant for herself)
- Focus group discussion about the avatar with all participants

The study session was video- and audio-recorded for later analysis.

1.4 Technical setup

The evaluation took place online via Zoom. Each participant used their own computer at home. During the app and avatar evaluation, all participants stayed in the Zoom meeting, but muted their audio and turned off their webcam. The facilitator was available for questions during these testing phases.

2 FOCUS GROUP DISCUSSION

2.1 App

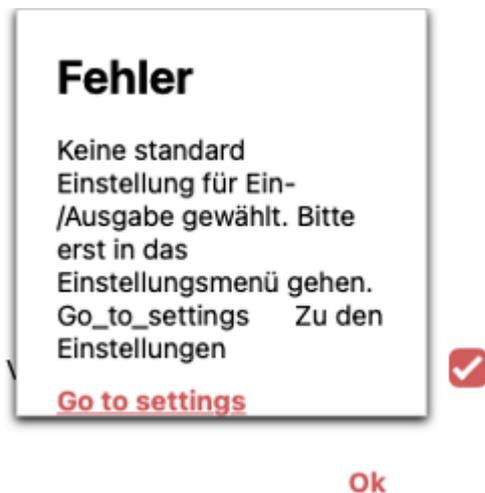
2.1.1 General feedback

The feedback was generally positive. However, since the translation module has not yet been implemented, the evaluation focused on the user experience only. This had the consequence that the full potential of the app could not be evaluated.

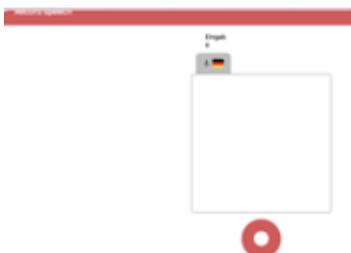
2.1.2 Handling and navigation

Study participants reported different issues regarding handling and navigation of the app. The possibility of changing to dark mode was perceived as something very positive. Some participants reported that the recording of the videos did not work and that there were also some issues with the flag icons and the language codes beneath them (e.g., the German flag is shown with “DSGS” beneath it; this should be “DGS”). It was also reported that not all icons were active at all times (e.g., recording of video).

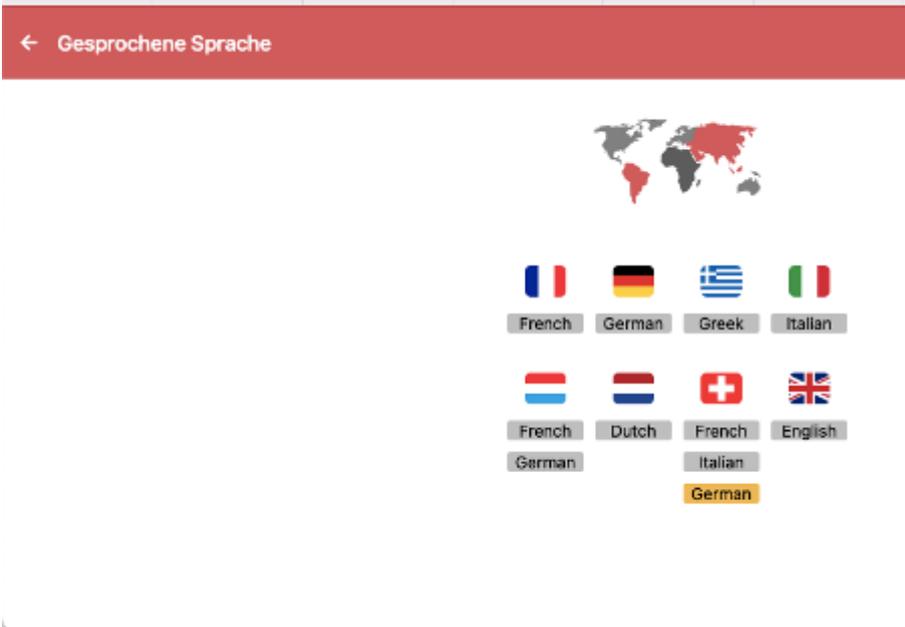
The participants provided screenshots of issues they encountered while testing the app, which are included below.



Comment by participant: “doesn’t work; remove checkmark”



Comment by participant: “not sure what is happening here”



Comment by participant: "Following this, nothing worked anymore"



Comment by participant: "Germany and Belgium have wrong labels 'DSGS'"



Comment by participant: “Buttons not active”

2.1.3 Recommendations and questions

The study participants suggested to make the navigation more user-friendly, for example, to move the arrow on the top left closer to the icons. Participants also did not find the *Settings* option immediately, only after a while (which should be more intuitive). Participants also recommended to use better/clearer icons, for example, for language modes (written, SL, spoken) or recording.



The participants further raised two questions:

- Is including Swiss German dialects as an option speech input/output planned?
- Is including speech into the app planned overall? Or only spoken language in written form?

2.2 Avatar

2.2.1 General feedback

The general feedback was neither positive nor negative, the focus was on the signed utterances of the avatar and the survey as a whole.

2.2.2 Linguistic issues

The participants discussed in detail the lack of non-manual components of the avatar, for example, raised eyebrows, no changes on the cheeks can be seen when the mouth moves (e.g., when signing VIELEN DANK), mouthing is often not clear to support comprehension

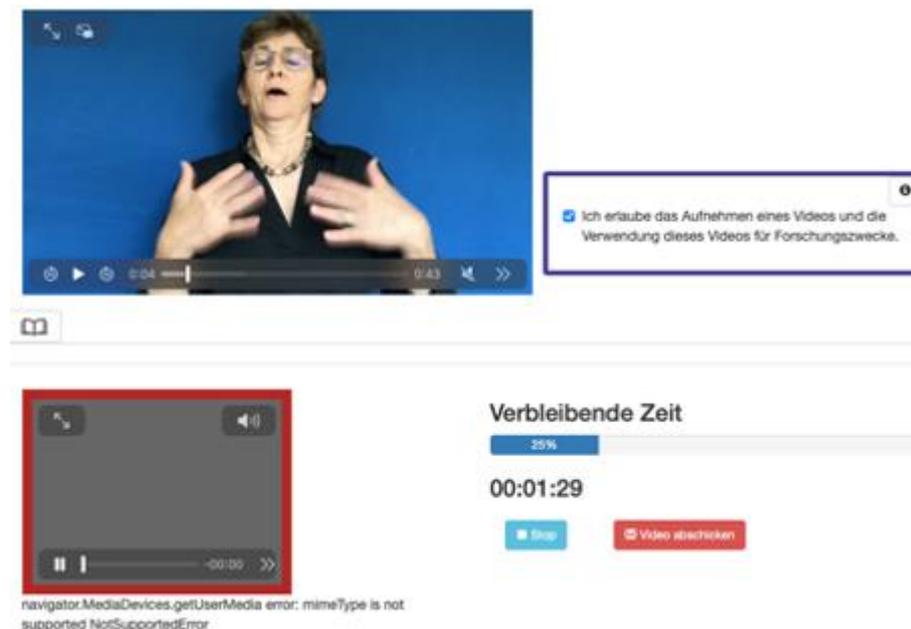
together with the manual components, the eyes should be wide open more often. Participants also mentioned that the movement of the upper torso did not look natural to them. In signs where one of the hands had contact with the other hand or another location at the signer's body, they deemed the location not always correct but still comprehensible.

Additionally, they remarked that the coordination of eyegaze when establishing/referring back to an index in space was often missing (i.e., eyegaze should move to the index when referring back to it).

Generally speaking, the participants commented that all productions of the avatar were understandable.

2.2.3 Feedback to survey

The study participants also provided some specific feedback regarding the survey. For example, the video format of the human signer was distorted. All participants reported that they were missing a return button and that they were not sure if the video recording worked. One participant suggested that an additional response option "only part of the sentence was understood" should be added. The participants reported that the instructions were very clear and good. They also observed that the avatar sometimes did not sign exactly the same as the human signer.



2.2.4 Recommendations

The participants' discussion centered around one major question (as opposed to a recommendation):

- Is the goal that an avatar should be as human as possible or that the content should be understood?

3 OTHER FEEDBACK

3.1 Feedback from facilitator (optional)

The evaluation of the app was challenging since the translation component was not yet implemented. This should be implemented for the final evaluation cycle. The procedure worked well.



FRENCH SIGN LANGUAGE DEAF GROUP

- **NAME OF PERSON WRITING REPORT: HENault-TESSIER MÉLANIE**
- **DATE OF EVALUATION: 22 OCTOBER 2022**
- **SIGN LANGUAGE: FRENCH SIGN LANGUAGE**
- **GROUP (DEAF/HEARING): HEARING**

1 METHOD AND PARTICIPANTS

1.1 Pilot study

A pilot study was conducted by Interpretis end of September in order to finalize the organisation of the two testing sessions. It was carried out with 2 participants: 1 hearing French/LSF interpreter and 1 signing deaf employee from Interpretis.

The pilot study showed that:

- ⇒ Participants did not know exactly what and how to evaluate the EASIER application so it was necessary to provide a guideline to help the participants in their observations and testing session.
- ⇒ The feedback and comments were extremely specific and concerned details of the application or the avatar.
- ⇒ Testing the app on a computer did not allow the participant to foresee its real uses and to evaluate it according to its intended purpose.

These findings led us to make two adjustments to make the testing sessions run more smoothly:

- ⇒ We clarified the elements to be considered and evaluated in the app and prepared a handout for the participants (Annexe I).
- ⇒ In order to facilitate feedback and comments as well as their understanding by all participants, we decided to use a video projector so that the users could precisely indicate the elements they were talking about.
- ⇒ We decided to test the app from mobile phones and asked participants to bring their own for the testing session.

1.2 Facilitator

Julie CLICQ was chosen as facilitator. She works for Interpretis as a client relations officer and is permanently in contact with deaf and hearing clients such as deaf associations and local institutions. She is herself a deaf signer and uses LSF.

Mélanie HÉNAULT-TESSIER, researcher at Interpretis, also took part in this evaluation session to help hosting the participants (breakfast, coffee, etc.) and with the technical set-up.

1.3 Participants

In order to constitute a group of deaf users, we emailed directly deaf persons belonging to the Toulouse signing community. We tried to vary the age and the profiles of the contacted persons. The email briefly introduced the EASIER project and described the testing session and related discussions. It also invited users to bring their own mobile phones in order to test the EASIER mobile application. Finally, we indicated that the participation to the testing session was not remunerated but entitled to a €50 voucher per person. The email was sent in both written and LSF versions.

We received four affirmative replies. However, a last-minute cancellation forced us to contact additional deaf people.

We finally managed to gather 4 deaf signers:

- 2 persons between 10 and 20 years old: one female, university art student; one male high school student.
- 2 persons between 46 and 60 years old: one female, e-commerce website manager; one male, deaf mediator.

Prior to the testing session, a consent document in both written and video LSF version was sent to participants. It described the EASIER project, explained the terms and conditions of their participation, asked for their agreement to the video recording and reminded them about their right of withdrawal.

1.4 Procedure

In order to ensure that testing sessions with the deaf users and the hearing French/LSF interpreters were conducted in a similar manner, we established a procedure to be followed. The testing session with the deaf users group was conducted as follows:

Welcome and thanks

- Welcome
- Presentation of Julie CLICQ and Mélanie HENAULT-TESSIER
- Participants' presentation
- Acknowledgements

Consent

- Presentation of the consent form, recall of participants' rights and participation modalities
- Signature of the consent form by participants

Voucher

- Distribution of the 50 € vouchers
- Signature of the document acknowledging reception of the voucher

EASIER presentation

- Presentation of the "Easier" sign
- Presentation of the project (consortium, languages, technologies, goals)

Description of the testing procedures

- 1st test: mobile app
 - To be done on mobile phone.
 - <https://easier-integration.nuromedia.com/>
 - Log in with user@example.com / user
 - Handing out the evaluation guidelines to participants (Annexe I)
- 1st focus group: mobile app
 - Participants' general impression of the app
 - Guided discussion on the application's features
 - Use of overhead projector so that participants can precisely refer to the elements they are talking about

Coffee break

- 2nd test: online questionnaire on the signing avatar
 - To be done on computer
 - <https://sign.ilsp.gr/slt-eval/>
 - There are no instructions, please follow the questionnaire
- 2nd focus group: signing avatar
 - Participants' general impression on the avatar
 - General questions:
 - Is the avatar intelligible?



- Does the avatar sign like a human?
- o Precise questions on the expressions and gestures of the avatar based on the videos in the questionnaire

1.5 Duration

The testing and discussion session lasted 3 hours and 45 minutes:

General presentation: 25 minutes

Application testing: 30 minutes

Discussion about the application: 1h20

Break: 20 minutes

Questionnaire on the avatar: 25 minutes

Discussion on the avatar: 35 minutes

1.6 Technical setup

The testing session took place in a large room in the Interpretis office. All participants were on site. The tables were set up in an "L" shape so that all participants could see each other, the facilitator as well as the wall on which the different visual elements were displayed.

As requested, participants all brought their mobile phones to test the app.

Four laptops were set up in the room so that each user could complete the online questionnaire.

An overhead projector was also installed so that we could project the user interface and the videos of the avatar to facilitate feedback and discussions.

We recorded the discussions with a single high quality camera with a wide angle lens.

The anonymised picture below shows the human and technical set-up.

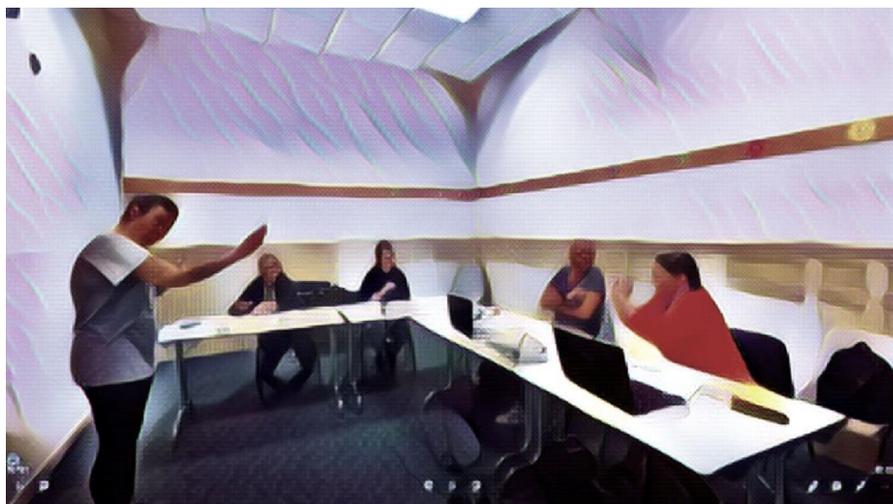


FIGURE 62- HUMAN AND TECHNICAL SET-UP

2 FOCUS GROUP DISCUSSION

2.1 EASIER APPLICATION - GENERAL FEEDBACK

2.1.1 General feedback

Comments on the EASIER application were generally unfavourable. Two criticisms stood out:

- The first general and spontaneous feedback from deaf users is that the application's interface is "*empty*" both in the information it delivers about how it works and its design. Consequently, deaf users say that its use is unintuitive (cf. 2.1.2) and the app seems unattractive (cf. 2.1.3).
- The second recurrent comment made by all deaf users concerns the interface's progression, which is not straightforward. There is therefore a sense of back and forth which creates confusion and does not allow for fluidity (cf. 2.1.4).

These feedbacks are more closely described and analysed in the sections below (2.1.2; 2.1.3; 2.1.4). We will extend this analysis by looking at the way the actual progression through the interface frames its actual use (2.1.5).

In the section 2.2, we present the deaf users' in-depth comments, organized by topic: terminology, the application various functionalities, and visual components.

2.1.2 Unclear use and usefulness

Deaf users consistently said that the application did not feel very intuitive to them: they had both difficulty in understanding what it could be used for and how to use it.

Users observed that the app's front page provides little information about its use and usefulness. More specifically, they remark that the EASIER logo gives little indication of potential uses. This feedback from users shows that the affordance of the application, i.e. the ability of an object or system to evoke its use and function, is not sufficient.

Furthermore, users reported that the handling of the app was not intuitive and required many clicks around to find out how it works.

2.1.3 Design & Aesthetics

Users note that there is "*a lot of white space*" and that the application's interface is not visually stimulating. As a result, users consider the application to be "*cold*" and mentioned that it fails to "*draw them in*".

A young deaf user, who is very familiar with the latest technology, pointed out that the interface is not "web responsive", which is one of the reasons why there is so much white space on some mobile phones. This creates another more important issue: the information at the bottom of the screen, especially the "Start" button, is hidden and requires the user to scroll up to catch the hidden information.

2.1.4 Progression within the app

Deaf users were very critical of the progression throughout the interface. They complained about the fact that it was not linear and requires constant steps backwards, particularly when defining the input and output communication modes and languages.

As shown in FIGURE 2, after selecting the input or output language the next page is not automatically displayed as users would like. Instead, the user needs to go back to the "Start" menu by clicking on the arrow in the upper left corner.

Users find these steps back "*painful*" and expect to be automatically driven to the next step.

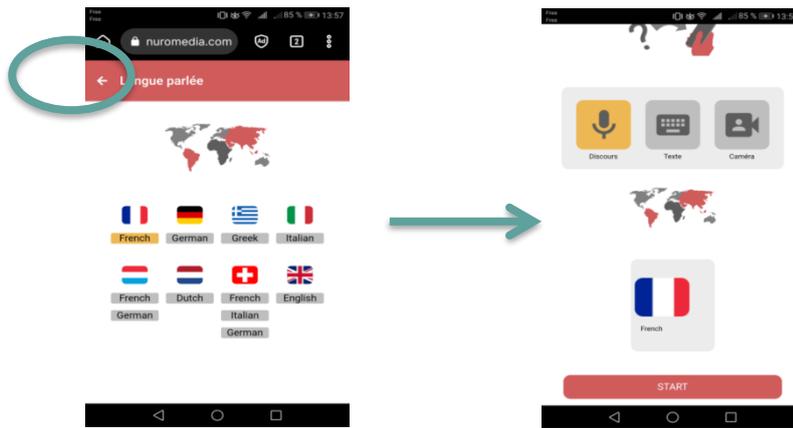


FIGURE 63-SELECTION OF THE INPUT AND OUTPUT LANGUAGES

2.1.5 Progression within the app and interaction flow

Deaf users have identified that the way the app is currently constructed constrains the shape of the interactions between the interlocutors and does not allow a smooth dialogue. On the contrary, it produces an interaction that feels like a consecutive translation instead of simultaneous one like real interpreters do.

The progression through the interface implies that the user first sets the input mode and language. This first action leads directly to the message recording. Once the message has been recorded, the user has to set the output mode and language to be able to generate the translation. The translation is then shown to his/her interlocutor who goes through these steps again to give his/her answer. The progression can be visualised as follows:

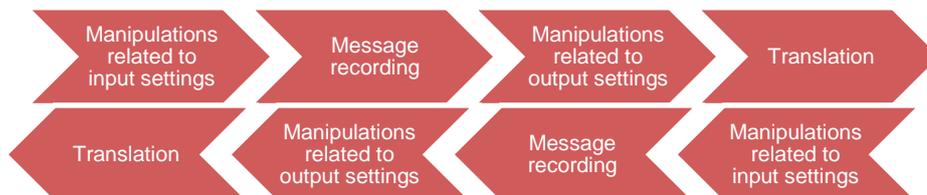


FIGURE 64-DISRUPTION OF THE SEQUENTIAL ORGANISATION OF TALK



The above diagram clearly shows that the natural sequential organisation of talk, based on adjacency pairs¹⁴, is disrupted by manipulations related to input/output settings.

Deaf users mentioned that these steps increase the sequentiality of the interaction, interfere with fluidity and remind them of conversations written on pieces of paper.

2.2 EASIER APPLICATION. IN-DEPTH REVIEW

2.2.1 English terminology in the French app

Deaf users have unanimously criticised the use of English words in the French version of the app, more particularly in the language dropdown menu (FIGURE 4), in the input and output setting pages and the "Start" button.



FIGURE 65-ENGLISH TERMINOLOGY

2.2.2 Favourite settings

Deaf users did not spontaneously understand what "Use favourite settings" proposed on the application's homepage referred to. They did not understand whether it referred to the communication mode and language choices or also to the avatar settings or the contrast settings.

Furthermore, as it can be seen on FIGURE 5, one user mentioned to have been confused by the double "settings" functionality. He said he did not understand right away that he first had to go into the settings in the drop-down menu in order to be able to select the "favourite settings" option afterward.

During the focus group, users talked a lot about the other applications they are used to and pointed out that these applications usually link their communication preferences to the "Profile" settings under the section "My Profile" or "My Communication Profile". As they are used to this, deaf users felt that "Settings" should be replaced by "Set Profile", which would make this feature more meaningful to them.



FIGURE 66-SETTINGS

They would also like to be able to save multiple profiles to quickly adapt to different situations, as they find the whole settings process time-consuming and complicated.

2.2.3 Input and output settings

When the user selects an input or output language, the selected language is highlighted in yellow as it can be seen on FIGURE 6. For two deaf users, this is not enough. They would

¹⁴ Schegloff, E. A., & Sacks, H. (1973). Opening Up Closings. *Semiotica*, 8, 289-327.

like a more visible validation of their choice and suggest adding a "save" icon at the bottom of the screen.

Moreover, deaf users do not like the dissociation of input and output default settings as shown in FIGURE 7. They would like to have the input and output choices presented on a single page, similar to the app Google translate. This proposal made by one participant was unanimously supported by all the others.

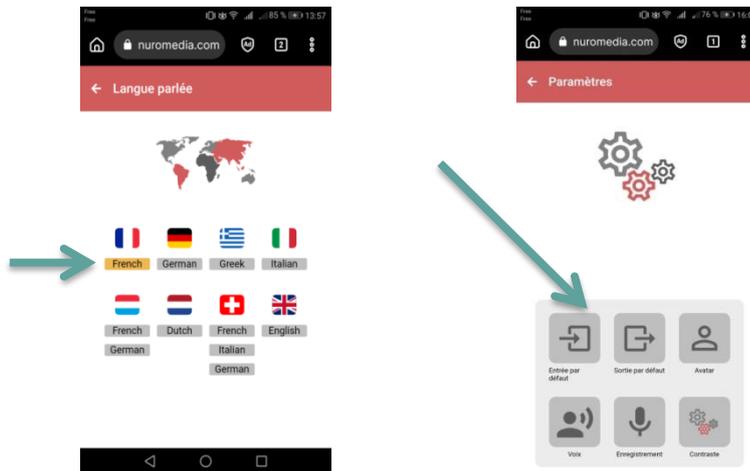


FIGURE 67- SELECTED LANGUAGE

FIGURE 68-INPUT & OUTPUT ICONS

2.2.4 Voice settings

This setting was not well understood by the participants who did not immediately understand the notion of "vocal style". Three comments were made by users about this setting:

- For this to be properly understood, the term "voice style" should be replaced by "language level".
- This setting should be proposed when choosing the input and output language.
- One user asked if this setting also applies to the sign language of the avatar (question we were unable to answer).

2.2.5 Contrast settings and font size

One of the deaf users was colour blind and had significant trouble seeing the pictograms which were not sufficiently contrasted for his needs. It turns out that he had not spotted the "Contrast" feature which he felt should be more easily noticeable.

In addition, when the contrast mode is activated on the app, the chosen language is no longer visible in the drop-down menu as showed in FIGURE 8.

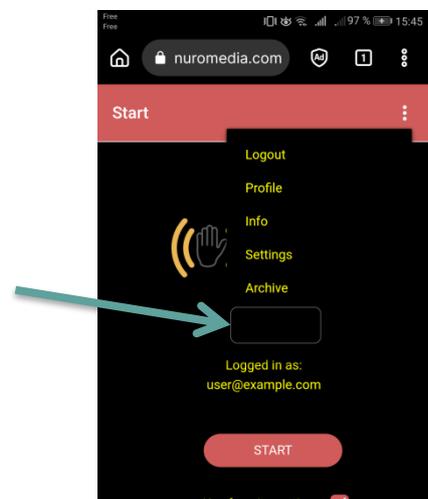


FIGURE 69-CONTRAST SETTING HIDING THE LANGUAGE PREFERENCE

2.2.6 Accessibility settings

Deaf users noticed that the app did not offer the usual accessibility settings usually found on the top left or right corner of any application or website.

2.2.7 Feedback on visual elements and terminology

	<p>This pictogram was not well understood by participants:</p> <ul style="list-style-type: none"> - The question mark is too vague and it is not clear whether to understand "who is speaking?", "how is he speaking?" or "who is he speaking to?" - It is quite similar to the output setting pictogram, which means that users have to concentrate to recognize the pictogram and make sense of it.
	<p>Deaf users found this pictogram confusing as they all tried to click on it and select the country on the map. Two of them also wondered why some countries were shown in red or dark grey.</p>
	<p>Selection of the spoken or written language :</p> <p>Deaf users indicated that country names should be added as they did not recognise all flags.</p>

	<p>SL selection:</p> <p>Users indicated that country names should be added as they did not recognise all flags.</p> <p>One user would have liked to see the dedicated sign language pictogram (two-hand pictogram).</p> <p>When choosing the written language for the translation output, the app suggests sign languages, which is a mistake.</p> <p>On the other hand, when choosing the language of the avatar, the app suggests vocal language, which is also a mistake.</p>
	<p>For more consistency, one user mentioned that the same pictogram should be used all over the app for vocal communication and voice recording, i.e. the microphone.</p>
	<p>Same as mentioned above for the input pictogram.</p>
	<p>In the "Settings" section, users did not understand the difference between these two pictograms.</p>

<p>Default input</p> 	<p>The meaning of this pictogram is not clear; users do not understand it until they <i>click</i> on it.</p>
<p>Default output</p> 	<p>Same as mentioned above.</p>
	<p>Users did not understand the word "<i>intonation</i>" as the following settings seems to refer to the language level rather than prosody.</p> <p>The pictograms were also misinterpreted as the blazer was associated with a male voice and the t-shirt with a female voice. The second row of parameters therefore appeared to be redundant.</p>
	<p>The "X" has been unanimously criticised because it is either not understood or considered to be deprecative.</p> <p>All users said that "Sexe" should be replaced by "Genre".</p>

	<p>The "contrast" feature was not understood as it seems to be redundant with the other features.</p>
	<p>The saving, copying and sharing features were not understood.</p> <p>Users asked whether it will be possible to record entire conversations or just bits of it.</p> <p>One user was also concerned about the security of archived or shared data.</p>

2.2.8 Recommendations

- Allow multiple profiles saving.
- Create a "Save" icon to validate language or profile choices.
- Default input and output settings to presented and choose on one single page (like Google Translate).
- Language level should be decided at the same time as the language choice.
- Language level should also be proposed for the avatar.
- Use the ad hoc pictogram to designate non-gendered individuals.
- Terminology should be entirely in French.
- Terminological changes:
 - Replace "Settings" by "Set Profile".
 - Replace "voice style" by "language level".
 - "Sexe" should be replaced by "Genre".

2.3 Avatar

2.3.1 General feedback

To briefly describe the content of the focus group on the avatar with the deaf users, we can say that they find that its intelligibility has improved but that it still does not sign like a human.

Deaf users acknowledge that the avatar has improved over the years and that its sign language has become more subtle. Nevertheless, all deaf users complained about its lack of expressiveness. The following comment made by one of the participants is an efficient summary of what the participants expressed during the focus group:

"It's not bad, the avatar has evolved. It has become more refined, but the work on facial expressions is not finished. The avatar's face is still empty." (DU4)

Regarding the lack of expressiveness of the avatar, two types of response could be observed during the focus group. The two youngest users, who both play video games, consider its expressiveness to be insufficient but consider that the avatar will evolve and eventually look like the characters in their video games. They consider its current state to be temporary so it does not cause any rejection. The two older users were more sceptical, one of them rejecting the technology quite radically:

"Everything was difficult to understand for me, I'm not happy with this solution. There is no humanity in the avatar, it's totally robotic." (DU3)

Besides the fact that its lack of expressivity leads to a robotic feeling, users also mentioned other consequences that we will explore in the sections below. We will also address the issues of body movements, rhythm and accuracy of the avatar's signs and examine how they contribute to the understanding of the message and the avatar's acceptance.

2.3.2 Facial expressions & mouthing

The deaf users noticed all facial expressions of the avatar: frowns, blinks, lip movements and, by extension, mouthing.

However, they would like to see these **facial expressions amplified** so that they can be more easily perceived. Indeed, since these expressions are very important for SL speakers, they should be prominent enough to be picked up without having to scan the avatar's face, as one of the deaf users mentioned it:

"Yes, there are expressions. The frown, when I take a detailed look, I see it. But when I keep an overview, I don't see the facial expressions; I have to look carefully to see them." (DU3)

Since sign language speakers generally adopt an overview rather than a focused gaze, they find it **less tiring** when facial expressions are amplified.

These facial expressions also **contribute to the understanding of the message** and hinder understanding when they are not clearly visible, particularly when signs have a similar configuration:

"The sign language of the avatar is not bad but it misses expressions, I did not perceive the difference between "hello" and "thank you"." (DU2)

The lack of expressiveness also creates an impression of **coldness** that acts as a sort of repellent despite the fact that the intelligibility of the avatar has improved as shown in the short extract below:

- *She says goodbye with a cold face, it really bothers me.* (DU4)
- *Yes, but we understand her clearly.* (DU1)

This impression of coldness goes hand in hand with the impression of an **impersonal sign language**, cleared of all the individual specificities that deaf people enjoy:

"Also, it has to be said that every person has its peculiarities, but not the avatar, it is empty, it has no specificities [...] there is nothing that warms up its expression." (DU2).

Lip reading is also a resource that deaf people rely on to understand the signed message. According to the focus group participants, the avatar was not mouthing enough as one of the participants mentioned it: *"I grew up speaking. I became a signer later on. So I also need to lip-read, but there, on the avatar's face, there was nothing, it was empty"*. (DU3)

In order to distinguish between the signs *"hello"* and *"thank you"*, users looked at the avatar's lips but this did not help them at all. Indeed, as one user remarked: *"the avatar is mouthing 'Hello', it's in English!"* (DU1). These comments show that the **mouthing** should not be an approximation and should **match the spoken language** of the user's country.

2.3.3 Body movements

The head, shoulder and torso movements were well identified and valued by deaf users who found them important as they also contribute to the understanding of the message: *"The shoulder movement is there, it's very positive! It's good!"* (DU4).

2.3.4 Rhythm

The discussions revealed that the **intelligibility of the avatar also relied on the rhythm** of its sign language: **a slow pace generating misunderstandings**:

"All the gestures are broken down, we see all the movements. It's misleading because my eyes are drawn to movements that I don't usually see. [...] It's a question of rhythm. The pace is slow and it gives significance to things that are not important". (DU4)

When the transition between two signs is too slow, the user's attention is drawn to secondary elements, such as the movement of the "supporting" hand or the motion induced by the succession of two signs, which disturbs the understanding of the meaning.

To illustrate this point, we can go back to the verb *understand* used by the avatar in one of the video of the questionnaire. This sign is usually made at the forehead (element 1, FIGURE 9), but seemed to be stretched outwards in the avatar's video. Deaf users were confused by a "false intermediate sign" (element 2, FIGURE 9) resulting from its very slow chaining to the negation sign "not" (element 3, FIGURE 9)



FIGURE 70- SLOW TRANSITION DRAW THE ATTENTION ON SECONDARY ELEMENTS

Given these issues, one user pointed out that it would be important *"to be able to adjust the speed of the avatar to either slow it down or speed it up."* (DU2)

2.3.5 Signs accuracy

The focus group enabled us to confirm the fact that deaf sign language speakers are very attentive to the accuracy of the signs made by the avatar and that each of the SL parameters – hand configuration, movements, placement, and orientation - is taken into account. Thus, a wrong hand orientation or the absence of repetition can lead to misunderstanding.

For example, the "wait" sign was misunderstood, as one user said, because the finger movement was not repeated (FIGURE 10): *"It's different, she doesn't sign "wait" the same way. She only does one hand snap. Usually there's a little repetition. It disturbed me because it made me think of "stop"."* (DU1)

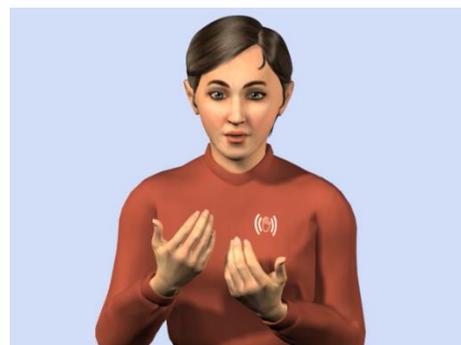


FIGURE 71- "WAIT"

The "sorry" sign (FIGURE 11) produced the same type of misunderstanding as shown by the following dialogue between two participants:

- There is a mistake in the sign, the hand is too tilted, and normally it should be flat. (DU2)
- And I make the sign on the palm instead. (DU4)
- Because of the hand's inclination, I first didn't understand anything. I understood when I saw it again. (DU2)
- Initially, I understood "there's no more". (DU4)



FIGURE 72- "SORRY"

2.3.6 Recommendations

- Amplify facial expressions.

- Mouthing should not be an approximation and should match the spoken language of the user's country.
- Create a speed setting functionality for the avatar.
- Have users validate the signs' accuracy.



3 OTHER FEEDBACK

3.1 Feedback from facilitator

The evaluation session and focus groups proceeded well and provided a lot of valuable insights. This seems to be a necessary step in any evaluation protocol. The focus group brought some granularity to the results of the on-line questionnaire and highlighted the elements to be improved in the avatar. Despite the fact that they are more cumbersome to organise, we nevertheless recommend to carry out focus group again during the next evaluation.

From our current experience, we can however propose some improvements:

During this evaluation, we noticed that deaf users did not consider the application as a tool for communication between two or more interlocutors. The application was understood more as a personal translation tool. We therefore recommend setting up evaluation sessions involving testing the application in an interaction situation between two interlocutors. In doing so, it would improve users' understanding of the potential uses of the application and would also link the analysis to concrete practices, which is currently not the case.

Concerning the avatar, we recommend that the evaluation focuses on whole sentences rather than on individual words, as we have observed misunderstandings related to the succession of signs. To a lesser extent, we would also point out that the answers to the question "where did you learn sign language?" in the online questionnaire often did not correspond to the users' experience. This section of the questionnaire should therefore be revised.

3 ANNEXE I

GUIDELINE FOR APPLICATION EVALUATION

- Progression through the interface: from entering the app to the production of a translation, how do you find the interface progression?
- Spatial organisation:
 - What is your opinion on the placement of information and visual elements on the interface?
 - Are the visual elements of the application clearly visible? Easy or difficult to find?
- Comprehension:
 - Instructions and actions: are the instructions/actions to be done understandable? (E.g. recording, translating, etc.)
 - Pictograms and logos: are the pictograms and logos understandable?
- Settings:
 - Are the settings easily accessible?
 - Are the settings useful? Are there any missing features? Are there any unnecessary settings?
- Terminology: do you think the lexical choices are relevant?
- Aesthetics: from an aesthetic point of view, how do you consider the application?
- Any free comments.





FRENCH SIGN LANGUAGE HEARING GROUP

- **NAME OF PERSON WRITING REPORT: HENault-TESSIER MÉLANIE**
- **DATE OF EVALUATION: 28 SEPTEMBER 2022**
- **SIGN LANGUAGE: FRENCH SIGN LANGUAGE**
- **GROUP (DEAF/HEARING): HEARING**

4 METHOD AND PARTICIPANTS

4.1 Pilot study

A pilot study was conducted by Interpretis end of September in order to finalize the organisation of the two testing sessions. It was carried out with 2 participants: 1 hearing French/LSF interpreter and 1 signing deaf employee from Interpretis.

The pilot study showed that:

- ⇒ Participants did not know exactly what and how to evaluate the EASIER application so it was necessary to provide a guideline to help the participants in their observations and testing session.
- ⇒ The feedback and comments were extremely specific and concerned details of the application or the avatar.
- ⇒ Testing the app on a computer did not allow the participant to foresee its real uses and to evaluate it according to its intended purpose.

These findings led us to make two adjustments to make the testing sessions run more smoothly:

- ⇒ We clarified the elements to be considered and evaluated in the app and prepared a handout for the participants (Annexe I).
- ⇒ In order to facilitate feedback and comments as well as their understanding by all participants, we decided to use a video projector so that the users could precisely indicate the elements they were talking about.
- ⇒ We decided to test the app from mobile phones and asked participants to bring their own for the testing session.

4.2 Facilitator

Mélanie HENAULT-TESSIER, sociologist and researcher at Interpretis, was the facilitator for the evaluation session with professionals French/LSF hearing interpreters.

Julie CLICQ, who works for Interpretis as a client relations officer, was also attending this testing session in order to prepare her own animation with the deaf users group.

4.3 Participants

It was decided to create a group of testers composed of sign language professionals whose profile best matched the profile of the interpreters likely to use the EASIER app. Indeed, the findings of the first study on intended uses and needs and of SL professionals carried out within the framework of the EASIER project (D1.1) showed that young interpreters were the most eager for translation tools. We therefore formed a relatively homogeneous group with:

- ⇒ 4 Interpretis French/LSF hearing interpreters.
- ⇒ All of them were junior interpreters who have worked as interpreters for 6 months to 5 years.
- ⇒ They were all women between 20 and 30 years old.
- ⇒ 3 of them knew about the EASIER project and 1 was discovering it.

Prior to the testing session, an email was sent to participants. It described the EASIER project, explained the terms and conditions of their participation, asked for their agreement to the video recording and reminded them about their right of withdrawal.

4.4 Procedure

In order to ensure that testing sessions with the deaf users and the hearing French/LSF interpreters were conducted in a similar manner, we established a procedure to be followed. The testing session with the interpreters group was conducted as follows:

Welcome and thanks

- Welcome
- Acknowledgements

Consent

- Presentation of the consent form, recall of participants' rights and participation modalities
- Signature of the consent form by participants

EASIER presentation

- Presentation of the "Easier" sign
- Presentation of the project (consortium, languages, technologies, goals)

Description of the testing procedures

- 1st test: mobile app
 - To be done on mobile phone.
 - <https://easier-integration.nuromedia.com/>
 - Log in with user@example.com / user
 - Handing out the evaluation guidelines to participants (Annexe I)
- 1st focus group: mobile app
 - Participants' general impression of the app
 - Guided discussion on the application's features
 - Use of overhead projector so that participants can precisely refer to the elements they are talking about
- 2nd test: online questionnaire on the signing avatar
 - To be done on computer
 - <https://sign.ilsp.gr/slt-eval/>
 - There are no instructions, please follow the questionnaire
- 2nd focus group: signing avatar
 - Participants' general impression on the avatar
 - General questions:
 - Is the avatar intelligible?
 - Does the avatar sign like a human?
 - Precise questions on the expressions and gestures of the avatar based on the videos in the questionnaire

4.5 Duration

The testing and discussion session lasted 3 hours and 30 minutes:

General presentation: 15 minutes

Application testing: 25 minutes

Discussion about the application: 1h02

Questionnaire on the avatar: 20 minutes

Discussion on the avatar: 1h05 minutes



4.6 Technical setup

The testing session took place in a large room in the Interpretis office. All participants were on site. The tables were set up in a "U" shape so that all participants could see each other and the facilitator. A French/LSF interpreter was standing beside the facilitator in order to allow Julie CLICQ to follow the voice conversations that were taking place.

As requested, participants all brought their mobile phones to test the app.

Four laptops were set up in the room so that each user could complete the online questionnaire.

An overhead projector was also installed so that we could project the user interface and the videos of the avatar to facilitate feedback and discussions. One interpreter had to change places during the discussions to see what was projected on the wall. We therefore opted for the "L" configuration for the group of deaf users.

We recorded the discussions with an audio recording device and took some pictures of the set-up.

The anonymised picture below shows the human and technical set-up.



FIGURE 73- HUMAN AND TECHNICAL SET-UP

5 FOCUS GROUP DISCUSSION

5.1 EASIER APPLICATION - GENERAL FEEDBACK

5.1.1 General feedback

Like deaf users, interpreters' comments on the app were generally unfavourable. Three criticisms stood out:

- The main comment made by all interpreters concerns the interface's progression, which is not straightforward. There is therefore a sense of back and forth which creates confusion, does not allow for fluidity and causes irritation (cf. 2.1.2).
- The second recurring feedback concerns the potential uses of the application which are not well understood by interpreters and which remain unclear even when manipulating the app (cf. 2.1.3).
- The third comment expressed repeatedly by the interpreters was that the application is unintuitive and requires a lot of click around to fully grasp its functionality (2.1.4).

These three general points are discussed in detail below (2.1.2, 2.1.3, 2.1.4).

In the section 2.2, we present the interpreters' in-depth comments, organized by topic: design, terminology, settings & functionality, and visual components of the app.

5.1.2 Progression within the app

Interpreters find the interface cumbersome and consider that it has too many pages. This feeling is reinforced by the way the progression through the application is organised. They complained about the fact that it was not linear and requires constant steps backwards, particularly when defining the input and output communication modes and languages.

As shown in FIGURE 2, after selecting the input or output language the next page is not automatically displayed as users would like. Instead, the user needs to go back to the "Start" menu by clicking on the arrow in the upper left corner.

Interpreters find these steps back "*painful*" and expect to be automatically driven to the next step.

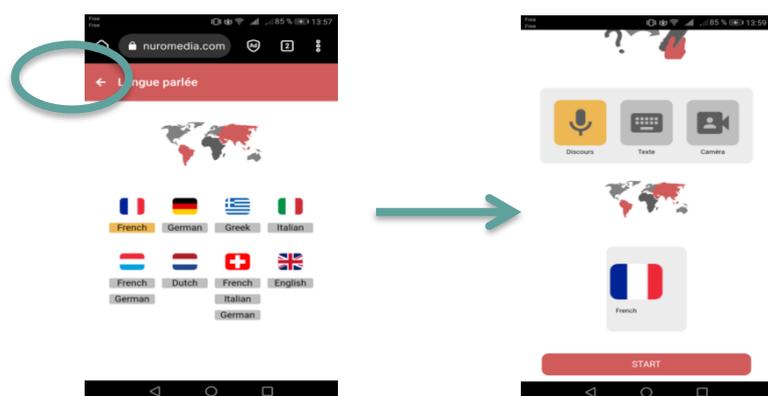


FIGURE 74-SELECTION OF THE INPUT AND OUTPUT LANGUAGES

Accessing the settings requires a return to the home page, which also hinders the progression. One interpreter also noted that the application did not offer to perform a second translation once the first was completed and that it was necessary to return to the start menu and repeat the entire process (FIGURE 3).

Finally, interpreters found that clicking the 'back' arrow of either their phone or the web page pulled them out of the app rather than going one page back; which resulted in some unforeseen back and forth activity.

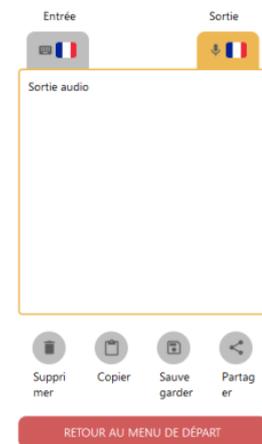


FIGURE 75-RETURN TO THE START MENU

5.1.3 Unclear intended uses

During the focus group with the interpreters, we were surprised to discover that the application was not perceived as a device to interact with another individual. This feedback shows that the affordance of the application, i.e. the ability of an object or system to evoke its use and function, is not sufficient.

Indeed, the application seems to be perceived as a personal translation tool. The copy/save/share functions contribute to this understanding and encourage interpreters to see it as a tool to create and share glossaries.

All interpreters pointed out the need to add information to clarify the potential uses of the app.

5.1.4 Unintuitive use

Like deaf users, interpreters consider the application to be unintuitive as it requires a lot of click around to find out how it works.

All interpreters pointed out the need to add information to explain how to use the app.

5.2 EASIER APPLICATION - IN-DEPTH REVIEW

5.2.1 Design & Aesthetics

One interpreter was really bothered by the layout. Like the deaf users, she found the white space too prominent and wondered why the design was not optimised. In her opinion, the various icons and pictograms could be enlarged for greater visual comfort.

The interpreters also pointed out that the interface was not "web responsive", which partly explain why there is so much with space on some mobile phones. They also reported that the "Start" button required the user to scroll up to be seen.

Interpreters generally considered the application to be a little too "sober" visually.

5.2.2 Saving input and output settings

When the user selects an input or output language, the selected language is highlighted in yellow as it can be seen on FIGURE 4. Interpreters would like a more explicit confirmation of their choice and suggest adding a "Save" icon at the bottom of the screen.



FIGURE 4- SELECTED LANGUAGE

5.2.3 Voice settings

Interpreters did not understand the word "*intonation*", related to prosody, as the following setting options rather seem to refer to the language level.

They didn't understand the option "Adapt intonation" either. They supposed that it referred to an automated adaptation tool using AI to match the translation to the vocal, written or signed style of the user.

Interpreters did not immediately understand the notion of "vocal style". As mentioned by the deaf users, interpreters also suggested replacing the expression "*vocal style*" by "language level".

The pictograms were also misinterpreted: the blazer was associated with a male voice and the t-shirt with a female voice. The second row of parameters therefore appeared to be redundant.

With regard to the possibility of choosing an English accent, interpreters wondered whether this setting would be proposed for other languages.

One interpreter wondered about the gender and the language level of the default voice. She stressed that it would be important to avoid the stereotype *male voice/sustained language level*.

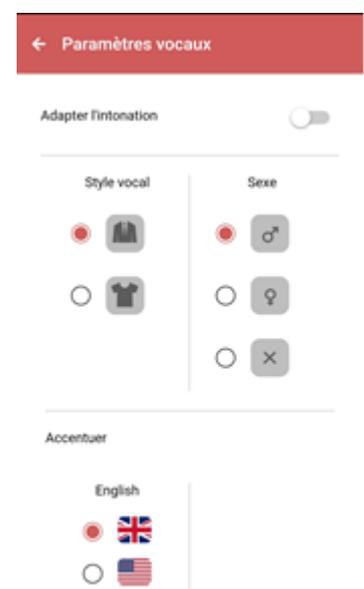


FIGURE 5- VOICE SETTINGS

5.2.4 Contrast settings and font size

The four interpreters tested the contrast mode and found several anomalies:

- When the contrast mode is activated on the app, the chosen language is no longer visible in the drop-down menu as showed in FIGURE 6.
- The drop-down menu for selecting a language appears as a white rectangle, which hinders accessibility (FIGURE 7).
- Input and output language choices are not visible because of the bright yellow thin font appearing on a grey background (FIGURE 8).
- Even in contrasting mode, the pictograms remain the same (dark grey/light grey), which does not make them very noticeable.

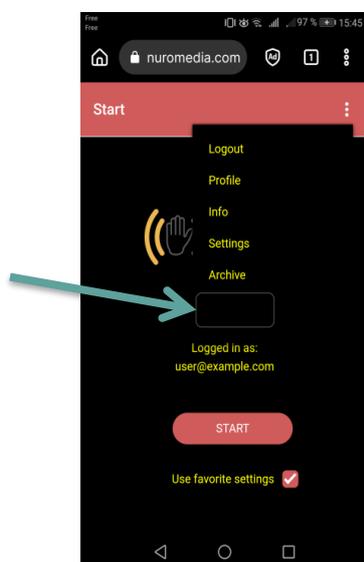


FIGURE 6-CONTRAST SETTING HIDING THE LANGUAGE PREFERENCE

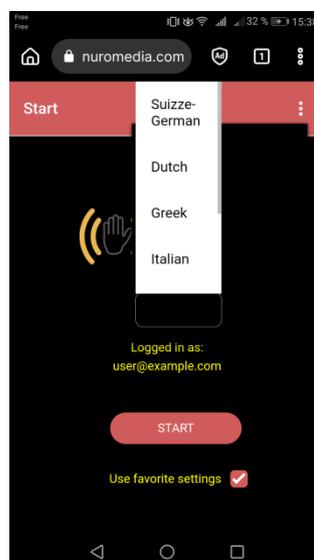


FIGURE 7-CONTRAST SETTING IS NOT APPLYING EVERYWHERE



FIGURE 8-LANGUAGE CHOICES WITH THE CONTRASTED MODE

5.2.5 Accessibility settings

The textual elements seemed too small to the interpreters. In conjunction with the possibility of increasing the font size, one interpreter therefore suggested adding a font thickening option.

More generally, interpreters noticed that the app did not offer the usual accessibility settings.

5.2.6 Terminology

Like deaf users, the interpreters noticed the use of English words in the French version of the app and asked for everything to be translated. This remark applies mainly to the language dropdown menu (FIGURE 8), the input and output setting pages, the “Next” and “Start” button.

The interpreters also identified literal translations from English that did not make sense in French:

- *Texte clair ? Oui/Non* (FIGURE 9). This expression should be replaced by: *Supprimer le texte ? Oui / Non.*

- *Accentuer* (FIGURE 10) should be replaced by “Accent”.



FIGURE 8-ENGLISH TERMINOLOGY



FIGURE 9-LITTERAL TRANSLATION "TEXTE CLAIR?"



FIGURE 10-LITTERAL TRANSLATION "ACCENTUER"

Interpreters also suggested small terminology changes in order to clarify the written information and help users find their way around the application:

- Change the name of the *Info* section concerning the project to *About EASIER*.
- The expression *Méthode d'entrée* is rather unusual; it could be replaced *Mode de saisie*.
- In the Input Method section, replace *Speech* by *Voice* under the microphone icon and *Camera* by *Sign Language* under the camera icon (FIGURE 11).
- On pages allowing selection of the input or output language, the term *Spoken language* appears in the top left-hand corner. This term is used even when selecting a written language or a sign language (FIGURE 12). Interpreters suggested to simply use *Language*, which suits all communication method, or adapt the name to the corresponding selection namely *vocal language*, *sign language and written language*.



FIGURE 11- INPUT METHOD



FIGURE 12-LANGUE PARLEE

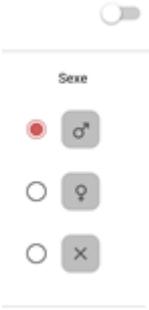
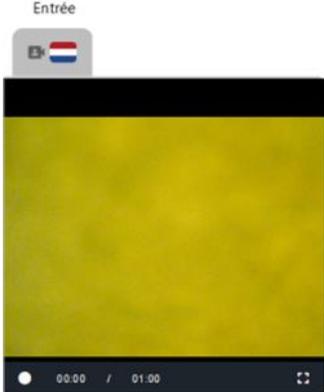
Finally, the interpreters noticed mismatches between the input method and the proposed language choices:

- When selecting the language for the camera input, the app only proposes vocal language. Sign languages are therefore missing.

- When choosing the written language for the translation output, the app suggests sign languages, which is a mistake.
- On the other hand, when choosing the avatar’s sign language, the app suggests vocal language, which is also a mistake.

5.2.7 Feedback on visual elements

	<p>This pictogram was not well understood by interpreters:</p> <ul style="list-style-type: none"> • As for deaf users, the question mark is too vague and can lead to several understandings.
	<p>Same as mentioned above for the input pictogram.</p>
	<p>Interpreters found this pictogram confusing as they all tried to click on it and select the country on the map. One of them also wondered why some countries were shown in red or dark grey and pointed out that the colours on the map did not match the languages supported by the application.</p>
	<p>Selection of the spoken or written language :</p> <ul style="list-style-type: none"> • Three out of four interpreters spontaneously clicked on the flag and were a little surprised to find that they had to click on the language name instead.
	<p>The “Recording” setting was not well understood by interpreters who wondered if this was another way of accessing the "Archives" file.</p>

	<p>One interpreter mentioned that the avatar pictogram was not saying anything about the avatar himself, and simply looked like a human pictogram. She would like to see a pictogram that clearly refers to an avatar. However, she understood it as the text underneath is speaking for himself.</p>
	<p>The "X" has been unanimously criticised by interpreters because it is considered to be depreciative.</p> <p>Interpreters unanimously said that "Sexe" should be replaced by "Genre".</p>
	<p>The "contrast" feature was not understood as it seems to be redundant with the other features.</p> <p>Interpreters would like to see the results of their current selection in order to verify it corresponds to their needs or habits.</p>
	<p>The record button on the camera device was difficult to find for one interpreter. She also mentioned that the sign language pictogram should be replacing the country flag.</p>

	<p>The saving, copying and sharing features stressed out many questions:</p> <p>Like deaf users, interpreters had many questions:</p> <ul style="list-style-type: none"> • Where will the files be saved? • Will there be a time limit for video recordings? • Why should I copy? Why sharing? • Can the saved data be deleted?
---	---

5.2.8 Recommendations

- Add information on how and for what purposes the EASIER app could be used.
- Provide access to the settings from all pages.
- Allow the production of several translations one after the other.
- Create a "Save" icon to validate language or profile choices.
- Language level should be decided at the same time as the language choice.
- Language level should also be proposed for the avatar.
- Use the ad hoc pictogram to designate non-gendered individuals.
- Terminology should be entirely in French.
- Terminological changes:
 - Replace *Text clair ?* par *Supprimer le texte ?*
 - Replace *Voice style* by *Language level*.
 - *Sexe* should be replaced by *Genre*.
 - Replace *Accentuer* by *Accent*.
 - Change the name of the *Info* to *About EASIER*.
 - Replace *Méthode d'entrée* by *Mode de saisie*.
 - Replace *Speech* by *Voice* under the microphone icon and *Camera* by *Sign Language* under the camera icon.
 - Replace *Spoken language* (that appears in the top left-hand corner when selecting a language by *Language* or *vocal language*, *sign language* and *written language*).
 - Correct the mismatches between the input method and the proposed language choices.

5.3 Avatar

5.3.1 General feedback

In general, interpreters were **positively surprised** by the quality of the avatar, which they found mostly intelligible and seemingly more refined than the other avatars they knew:

Actually, the rendering is quite good, I was expecting less quality in the signs (Int. 3)

I didn't expect it to be so clear (Int. 2)

I find it quite understandable on most videos (Int. 1)

In particular, the interpreters valued:

- The blinking;
- The thin fingers well separated from each other;
- The mobile phalanges;
- The shoulder and/or torso movements;
- The use of facial expressions

Nevertheless, as we will discuss in further detail below, although interpreters did perceive improvement in the avatar's expressivity, they unanimously stressed that the **avatar's facial expressions still needed to be accentuated**. One interpreter resumed it by saying that facial expressions *still needed to be taken up a notch* (Int. 4).

Another recurrent remark heard during the interpreter's focus group is that the avatar's sign language **pace was too slow**, resulting in a robotic appearance and possible misunderstandings:

I find that the gestures are a bit mechanical and the rhythm is very slow. The signs look rather detached from one and other; but it works anyway! (Int. 1)

5.3.2 Facial expressions & mouthing

The interpreters' criticisms of the avatar's expressivity are similar to those expressed by the deaf users: all of them acknowledge an improvement but still **consider it to be insufficient**:

It's the first time I look at an avatar so closely, but I find the facial expression very weak. Sometimes there is no eyebrow movement, there is nothing going on in his face. (Int. 4)

In the focus group, interpreters, like the deaf users, showed that they could not separate signs and facial expressions when evaluating the quality of the avatar's sign language. The lack of expressiveness of the avatar therefore **affected the intelligibility of the message**. It also **hindered the evaluation of the quality** of its signed expression. Finally, the lack of expressiveness of the avatar made it difficult for the interpreters to say that the avatar was signing like a human.

The discussions around the avatar's facial expressions also showed that the interpreters have **high expectations** concerning their subtleties and complexities. During the group viewing of the video of the avatar signing *Bonjour* (FIGURE 4), one interpreter mentioned that a small smile was missing on the avatar's face and that, as a result, the greeting seemed cold. A second interpreter then mentioned that the addition of a smile should imply a squinting of the eyes as these two expressions are linked. The video was then watched twice to determine if the eyebrows were mobile. As their movement was very gentle, the interpreters unanimously said that it would be important to **accentuate** it along with adding a smile and a slight crinkle at the eyes so that it can be considered as signing like a human. Similarly, the video of the avatar saying *Désolée* (FIGURE 5) was watched many times by

the panel of interpreters before its lower lip movement was noticed. All these examples plead for a generalised **amplification of the avatar's facial and body expressions**.



FIGURE 4 - BONJOUR (HELLO)



FIGURE 5 - DÉSOLÉE (SORRY)

To better appreciate these results, it is important to mention that interpreters answered the avatar questionnaire on a computer and that none of them watched the videos in full screen format. This means that the videos watched were about the size of a mobile phone screen. At this scale, some facial movements become barely perceptible to interpreters. This partly explains their wish to see these expressions amplified. Therefore, as the avatar's facial expressions were not very noticeable, **its expression was judge as quite neutral**, comparable to the standardised voice used in public transportation or GPS:

There is something of the automatic voice because of its barely marked facial expressions (Int. 1)

One last element concerning the avatar's facial expressions is interesting to be pointed out: interpreters noticed that the shape of the avatar's eyes and the location of her eyebrows (away from her eyes) gave her a surprised look. One interpreter stated that these physical details *gave an intonation to the speech, as if she was always in a state of surprise* (Int.2).

A small comment should also be made regarding mouthing as lip reading is also a resource for interpreters who rely on it to understand the signed message. Like the deaf users, interpreters scrutinized the avatar's mouth to distinguish between the signs "hello" and "thank you" done by the avatar, but this did not help them at all because the mouthing was in English. This shows that the **mouthing** should not be an approximation and should **match the spoken language** of the user's country.

5.3.3 Body movements

The head, shoulder and torso movements were well identified and valued by interpreters who found them **important** as they also contribute to the understanding of the message. Nevertheless, like facial expressions, interpreters mentioned that they wished they were **more prominent**. More specifically, the interpreters observed that the shoulders were not mobile when they should have been, for example when the avatar signs *sorry*. This sign is often accompanied by a shoulder shrug that interpreters expected to see on the avatar

This remark shows that the interpreters' expectations towards the avatar's sign language are the same as those they have towards any signing person.

5.3.4 Rhythm

Interpreters find the signed expression of the avatar much too slow, which was expressed in the following ways: *the rhythm is slow; I looked for a speed up button on the videos!* (Int. 1)

Interpreters therefore want to be able to speed up the videos as they wish and as they need.

Similarly to the deaf users, interpreters pointed out the intelligibility problem related to the sign language pace of the avatar (cf. section 2.3.4 of the Interim evaluation report_Deaf users by Interpretis).

Furthermore, the rhythm issue led to a discussion among interpreters about its significance and the fact that it should not be constant. Indeed, interpreters argued that **rhythm should evolve according to the sentences**, since it has the same role as punctuation in building the meaning of the message:

Rhythm plays the same role as punctuation. If there is a comma or a dot, we will mark micro-silences. In sign language, there are moments when we stop for a short time before moving on to the next part. These are rhythmic mechanisms which allow us to make sense of the text. (Int. 1)

5.3.5 Signs accuracy

Sign accuracy was an issue that consumed a significant part of the discussion between interpreters: incorrect hand configuration, improper motion, placement, or orientation of the hands resulted in substantial comprehension problems.

The sign that was most misunderstood was *Prêt* (FIGURE 6), which lacked its usual characteristics. It was missing the semi-circular and descending movement of the active hand and the 90° rotation of the supporting hand. The interpreters understood this sign as *Je me présente* (I am presenting myself), which was not congruent within the sentence. This example highlights the importance of following the precise parameters of each sign made by the avatar, as a wrong execution can jeopardize the intelligibility of an entire sentence.

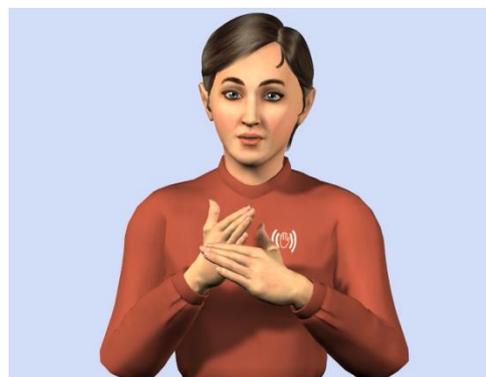


FIGURE 6 - PRÊT (READY)

5.3.6 Miscellaneous

One interpreter out of the four attending the focus group mentioned that the shadow cast by the avatar's hands and arms on her torso made her feel uncomfortable. She pointed out that interpreters and translators are usually lit so that there are no visible shadows.

5.3.7 Recommendations

- Amplify facial expressions.
- Add complexity in facial expressions, e.g.: smile>squinting > eyebrow movements.

- Mouthing should not be an approximation and should match the spoken language of the user's country.
- Create a speed setting functionality for the avatar.
- Use standard signs recognised by the official Deaf organisations of each country.



6 OTHER FEEDBACK

6.1 Feedback from facilitator

The evaluation session and focus groups proceeded well and provided a lot of valuable insights. This seems to be a necessary step in any evaluation protocol. The focus group brought some granularity to the results of the on-line questionnaire and highlighted the elements to be improved in the avatar. Despite the fact that they are more cumbersome to organise, we nevertheless recommend to carry out focus group again during the next evaluation.

From our current experience, we can however propose some improvements:

During this evaluation, we noticed that deaf users did not consider the application as a tool for communication between two or more interlocutors. The application was understood more as a personal translation tool. We therefore recommend setting up evaluation sessions involving testing the application in an interaction situation between two interlocutors. In doing so, it would improve users' understanding of the potential uses of the application and would also link the analysis to concrete practices, which is currently not the case.

Concerning the avatar, we recommend that the evaluation focuses on whole sentences rather than on individual words, as we have observed misunderstandings related to the succession of signs. To a lesser extent, we would also point out that the answers to the question "where did you learn sign language?" in the online questionnaire often did not correspond to the users' experience. This section of the questionnaire should therefore be revised.

4 ANNEXE I

GUIDELINE FOR APPLICATION EVALUATION

- Progression through the interface: from entering the app to the production of a translation, how do you find the interface progression?
- Spatial organisation:
 - What is your opinion on the placement of information and visual elements on the interface?
 - Are the visual elements of the application clearly visible? Easy or difficult to find?
- Comprehension:
 - Instructions and actions: are the instructions/actions to be done understandable? (E.g. recording, translating, etc.)
 - Pictograms and logos: are the pictograms and logos understandable?
- Settings:
 - Are the settings easily accessible?
 - Are the settings useful? Are there any missing features? Are there any unnecessary settings?
- Terminology: do you think the lexical choices are relevant?
- Aesthetics: from an aesthetic point of view, how do you consider the application?
- Any free comments.





DUTCH SIGN LANGUAGE DEAF GROUP

- 7.1 NAME OF PERSON WRITING REPORT: ONNO CRASBORN**
- 7.2 DATE OF EVALUATION: 2 NOVEMBER 2022**
- 7.3 SIGN LANGUAGE: NGT**
- 7.4 GROUP (DEAF/HEARING): DEAF**

1 METHOD AND PARTICIPANTS

1.1 Facilitator

- We ended up having Onno Crasborn (prof. of NGT at Radboud University, hearing) leading this group, because of the time pressure. Onno is a certified NGT interpreter and has been working with deaf team members since 1997.

1.2 Participants (as much as you can fill out, but still preserve anonymity)

- Two deaf research assistants who were unfamiliar with EASIER functioned as participants.

1.3 Procedure

- Participants met in one of our offices, read the info form and signed the consent form. Both worked on macs with large 27" screens. They used Safari to go to the url, registered and logged in, and played around. They chatted together as they went along, exploring all corners of the interface, but gradually getting confused and irritated with the fact that it was a dummy. Total evaluation lasted a little over an hour (but see below for the ensuing discussion). The moderator took quick notes and immediately created this report afterwards. No recordings were made.
- A comment at the outset was that the information form in Dutch was clearly translated by a Flemish person. This was just noticeable and didn't impede understanding, but it's something to take into account perhaps for the future, thinking about website and app translations. Not all deaf people being highly literate, small differences between language areas might have more impact than we think.

1.4 Technical setup

- See above.

5 FOCUS GROUP DISCUSSION

1.5 App

1.5.1.1 General feedback

The overall response was neutral, participants were open and curious as to what the app might do. In the end, participants were annoyed about the futility of the task, and the fact that their costly time went into these simple and 'obvious' observations. 'Is this academic level, this app development', one participant asked, shaking her head? When they heard that this evaluation was already done in multiple countries, they were both very surprised, not understanding why replication across different language groups would be necessary. The participants got into a long discussion about the ethics of it all, hearing that the evaluation of this phase of development would still be planned with hearing professionals (interpreters or teachers). In summary, they found it ethically irresponsible to do this, given the shortage of sign language teachers and interpreters in the Netherlands. Especially given the simplicity of the task. And wouldn't the same apply to other countries?! As a researcher, I can only concur, and I promised the participants that we would not carry out the evaluation with hearing sign language professionals for that reason.

Some smaller general observations:

One participant could imagine the interface is hard to use for someone with Usher's, impaired vision.

Starting with text input, nothing happens upon going to the next stage; a message saying 'this is just a dummy, translations not yet available' would have been nice.

Various setting options other than language preference were observed to be clear and simple, nice.

Themes addressed by comments below:

- A bug (1.5.2)
- Terminology and culture-related issues (1.5.3)
- Language options (1.5.4)
- Interface / Ease of use (1.5.5)

1.5.2 Bug

At some point, for one participant the list of translation languages appeared in some Scandinavian language after setting interface language to Dutch (names ending in -sk); couldn't reproduce this later on. On another computer, the language names were still in English, see screen shot.

← Gesproken taal



1.5.2.1 Terminology and culture-related issues

Activating the camera leads initially to an icon showing a microphone; this was noted to be confusing and not culture-appropriate.

The registration option 'Man/vrouw/non-binary' was very positively evaluated by one participant.

1.5.3 Language options

Participants did not find the interface language button, kept using it in English until moderator pointed it out after some ten minutes. One thought setting the preferred language would lead to a different interface (changing interface language not translation language). Even in the three-dot menu, the interface language is downplayed by the smaller font.

There was initial confusion about the languages available, where is NGT?

The language labels were found to be very small to click on, one would expect it would also be possible to click on the flags.

Participants kept selecting languages, their preference didn't seem to be stored. The presence of all the language options that are not selected are confusing, one found. One would have expected a 'preferred language' or 'default language' options in the Setting menu, but nothing there they thought. Thus, 'standaardinvoer' and 'standaarduitvoer' (standard input/output) are not self-explanatory enough, both users weren't attracted to try out what's behind there. One user specifically asked the moderator 'what does that mean?'.

1.5.4 Interface / ease of use

Registration was confusing, as the moderators instructions initially were to use the dummy login, which didn't work.

Participants found the interface not to be very intuitive after first login, some step-by-step guidance upon first login would be appreciated. "What am I supposed to do now?"

Saving settings with an explicit button was expected by one participant, instead, just navigating back seems to save the settings?

All the navigating back to the previous screen with the small arrow at the top left is not very user-friendly, both participants commented. The button is too small, and it is in an inconvenient location.

1.5.5 Recommendations

- Did participants mention any specific recommendations for improvement?
 - of the evaluation design?
 - of the app?
 - of use-cases?

1.5.6 Key quotes

- "I would have expected a higher quality level if this is made by or for a university."



2 OTHER FEEDBACK

2.1 Feedback from facilitator (optional)

- Please make the final evaluation more targeted and efficient, not distributing it over too many language groups, and critically evaluating whether it is really necessary to call on hearing language professionals given that they are sorely needed in other areas of society.



DUTCH SIGN LANGUAGE HEARING GROUP

- 7.5 **NAME OF PERSON WRITING REPORT: HOPE E. MORGAN**
- 7.6 **DATE OF EVALUATION: NOVEMBER 28, 2022**
- 7.7 **SIGN LANGUAGE: NGT (BUT NOT APPLICABLE)**
- 7.8 **GROUP (DEAF/HEARING): HEARING**

1 METHOD AND PARTICIPANTS

1.1 Facilitator

- Name of facilitator: **Hope E. Morgan**
- Brief description of facilitator: **employee of Radboud University; hearing signer (but not a signer of NGT); English-dominant speaker; linguist (Ph.D.; sign language focus)**

1.2 Participants

- How many participants?: **4**
- How were they recruited?: **Recruited through email with people who have worked on sign language linguistics projects before.**
- Describe the group with respect to age, gender, background/profile: **three women, one man; ages are 30s-60s; all are hearing, Dutch-dominant, English-fluent, NGT signers**
- Was there anyone else present in the session (other than participants and facilitator)? **No.**

1.3 Procedure

- Which components were tested? The
- How was the session organised? **It was a 1hr Zoom session (no one in the same room), conducted in English. First, participants were welcomed. There were some technical difficulties accessing materials that took a few minutes to work out; while that was being figured out, participants were asked to create user accounts and log into the app. Next, a consent form was emailed to everyone. Then, the introduction to the project was read in English, followed by the introduction to the app, also read in English (but was translated by GoogleTranslate to English from Dutch). Then participants were given 10 minutes to explore the interface. After that, we came back together**
- What was the order of components? **Registration; default settings; interface language; visual layout; archive**
- How long did each individual component take? **Around 5 minutes each, with the most time spent on the default settings**
- What was the total duration of evaluation? **One hour**

1.4 Technical setup

- Was the evaluation online or in person? **Online**
- What equipment was used?
 - **Each participant used a computer to join the Zoom.**
 - **For the evaluation, these were the devices used for the app interface:**
 - **Participant 1: desktop computer (PC)**
 - **Participant 2: desktop computer (PC) and an Android phone**
 - **Participant 3: iPhone**
 - **Participant 4: laptop computer (Mac)**
 - **No recording was made**

-What did the setup look like? **Zoom screen with equal windows; i.e., neither pins nor screen sharing**



2 FOCUS GROUP DISCUSSION

2.1 App

2.1.1 General feedback

- Was the response generally positive, or negative? **Generally positive**
- What guidance did the facilitator give to the participants? **No guidance was given at all.**
- What did participants use to complete the questionnaire (mobile phone, laptop computer, etc)? **We did not complete a questionnaire.**
- Which things did participants like, which did they not like?
- Introduce major themes that came up (can add more than 3 Theme sections)

2.1.2 Theme 1: Good visual experience with helpful choices

The visual layout got positive response for being simple and clean. The icons were generally transparent (clearly interpreted) and fit the meaning; e.g., one person specifically liked the hand holding a phone, with arrow coming out to a question mark. The number of choices available for changing the contrast, text size, and avatar appearance were appreciated.

One text-based comment: “Nice that one can see not just the translation output but also the input. On a laptop browser, plenty of space to present those side by side like in Google translate. Having to click the (small) tabs is unnecessary extra clickwork.”

A couple recommendations were made on this theme:

- Have different skin tone choices for avatar
- In Dark Mode (Mac desktop, Safari), the language labels are yellow text on light grey background, which is difficult to read; improve contrast?
- In avatar settings, be able to preview changes to avatar as different radio buttons are selected?
- The word “contrast” in the Avatar settings is a bit confusing; what does it change?
- Be able to click the flag icon as a button when there is only one language choice
- Need clarity about the Luxembourg flag: is this the correct flag and language combination?

2.1.3 Theme 2: Where am I in the process?

In general, the “wayfinding” through the app was found to be problematic. The use of so many back arrows led to confusion about the status of the choices that were made, and what point in the process that a user is at, in terms of choosing initial settings versus choosing session-specific settings versus providing input versus expecting output. Each of these steps/phases are not indicated in an intuitive way. One participant noted, “the click dummy is making us click (a lot),” but that it isn’t linked to a clear logical progression or some type of menu structure.

Most participants agreed that some type of confirmation is needed once a setting choice has been made (at all levels), though one person using the mobile app (iPhone) thought that it was relatively clear because of color changes she noticed.

Also, the order and access to default settings could be improved. For example, a participant said that app users are familiar with an initial screen to establish the language of the

interface (typically using a flag), but the order that it is done here doesn't feel right. Also, confrontation with an error message to set up the defaults "almost feels rude". Half of the participants didn't notice the default setting option in the menu (possibly desktop users only), while in the mobile app, the menu with the settings and other choices was actually too prominent.

2.1.4 Theme 3: Which languages?

As a lesser theme, there were some general issues about language selection. Near the end, one person asked about selecting a sign language; they had never seen the sign languages as an input or output option. Also, as mentioned in Recommendations below, the options for languages were not always clear in different ways.

2.1.5 Recommendations

-Registration

- One person said the verification during registration felt like one step too many, but other participants thought it was fine

-Default settings

- After making Dutch the default interface language, if you go back to change the interface language (in pop-up menu), the language choices are not Dutch! They all end with -sk, so it may be a Scandinavian language? (e.g., Englesk, Tysk, Fransk)
- Should be a clearer path to access these settings at any time
- See avatar setting suggestions mentioned in Theme 1 above
- Something is odd about having the "favorite settings" checkbox on the first screen as well as default options?
- None of participants understood what "adapt intonation" meant
- The icons for voice style were not clear for one participant (why clothes?)
- Selection of English style:
 - Why is American English an option for a Europe-based app?,
 - Even when choosing a different default/interface language (i.e., Dutch), the British/American option is still shown; why not shown variants of Dutch?
 - Even if you select "American English", the app still says "British"

-Archive settings

- Missing options/functionality:
 - *review/playback* feature
 - *save* option
 - *share* option
 - *delete* option

-Translate settings

- Why is there no Belgian flag, but there is a Luxembourg flag?
- Why can't you click on the flag? It feels awkward not to be able to click on it
- Participants wondered: is 'German' the same in different countries?



3 OTHER FEEDBACK

3.1 Feedback from facilitator (optional)

-How did the procedure go?

- It was fine, other than a little confusion in not being able to easily locate the actual link to the app (!)
- One hour was perfect; more would have not seemed worth it for just the app.

-What things could be changed for the final evaluation cycle?

- Make just one link or one PDF 'info sheet' for facilitators to find everything needed and in a very clearly organized way: link to app, instructions, materials. Then share that link or PDF in a very clear way with facilitators.





GREEK SIGN LANGUAGE DEAF GROUP

- 1. NAME OF PERSON WRITING REPORT:**
- 2. DATE OF EVALUATION:**
- 3. SIGN LANGUAGE:**
- 4. GROUP (DEAF/HEARING):**

1 METHOD AND PARTICIPANTS

1.1 Facilitator

- Name of facilitator: Kiki Vassilaki
- Brief description of facilitator
 - internal to organization
 - hearing coda
 - language background: GSL & Greek bilingual
 - professional background: GSL expert (PhD in Sign Linguistics), member of the Sign Language Technologies Team at ATHENA/ILSP

1.2 Participants

- How many participants?
Four (4) experts (+1 in the pilot) participated in the evaluation of both the EASIER avatar and the app design.
- How were they recruited?
Participants were recruited by invitation among collaborators in other SL projects of the Institute.
- Describe the group with respect to age, gender, background/profile
The group was composed by two female and three male deaf GSL signers (including the signer who performed the pilot). One participant was between 60 and 70 years old, working as an administrative secretary. One participant was between 50 and 60 years old, working as a sign language teacher, One participant was between 30and 40 years old, working as a professional photographer, and one participant was a 20 year old high school student.
- Was there anyone else present in the session (other than participants and facilitator)?

In the session both hearing and deaf group facilitators were present. However, the hearing group facilitator did not take active part in the procedure.



Figure 1: Technical staff controlling the wide view camera.

Furthermore, the two members of the technical staff controlling the cameras entered the session area once or twice when it was necessary to change the camera batteries or check camera condition, without disturbing the session.

1.3 Procedure

-Which components were tested?

Two components were tested: (i) the EASIER avatar, and (ii) the EASIER app design.

-How was the session organised?

First, participants were welcomed to the institute. Then we explained the background of the project. Then participants signed the related consent forms. Then we proceeded first with evaluation of the EASIER avatar and next with the EASIER app. The session was concluded with discussion among the evaluators and the facilitator on the user experience, the legibility of the avatar, the strong and weak points of the implementation so far and evaluators' suggestions as to what should be changed or improved during the next period in respect to the two evaluated components.

-What was the order of components?

Evaluators first viewed the EASIER avatar and then the app mock-ups. Similar to the experience of evaluation by hearing experts, this proved to be evaluation methodologically preferable, since they first got familiar with the SL representation engine of the project and its current state of development, so that they then could better follow the incorporation of the avatar in the app design.

-How long did each individual component take?

Introductory project presentation and consent form signing lasted about 15 minutes. The evaluation of each component had a duration of approximately 20 minutes, while another 30 minutes lasted the overall discussion following both components evaluation. This made a total of 70 minutes of evaluation time. A 10 minutes break was made between evaluation of the avatar and the app. Another break of equal duration preceded the discussion session.

-What was the total duration of evaluation?

The overall duration of the evaluation was approximately 105 minutes

1.4 Technical setup

-Was the evaluation online or in person?

The evaluation was in person and took place at ATHENA/ILSP premises

-What equipment was used?

- For interacting with the components there were used the evaluators' smartphones, two institution laptops and two institution desktop devices.
- For recording participant discussion, three GoPro cameras were used mounted on the ceiling of the evaluation room, as well as one HD camera on a tripod to capture a global view of the room. All cameras were High Definition, recording in 1920*1080 pixels and 50fps (frames per second).
- Other equipment involved a projector, via which the facilitator projected the EASIER app during the app evaluation.

-What did the setup look like? (description, drawing or anonymised image of room + participants + facilitator + cameras)

The camera and discussion panel setup is depicted in the following figures.



Figure 2: Evaluation room with GoPro setup. L indicates left side camera, C indicates central camera, R indicates right side camera.

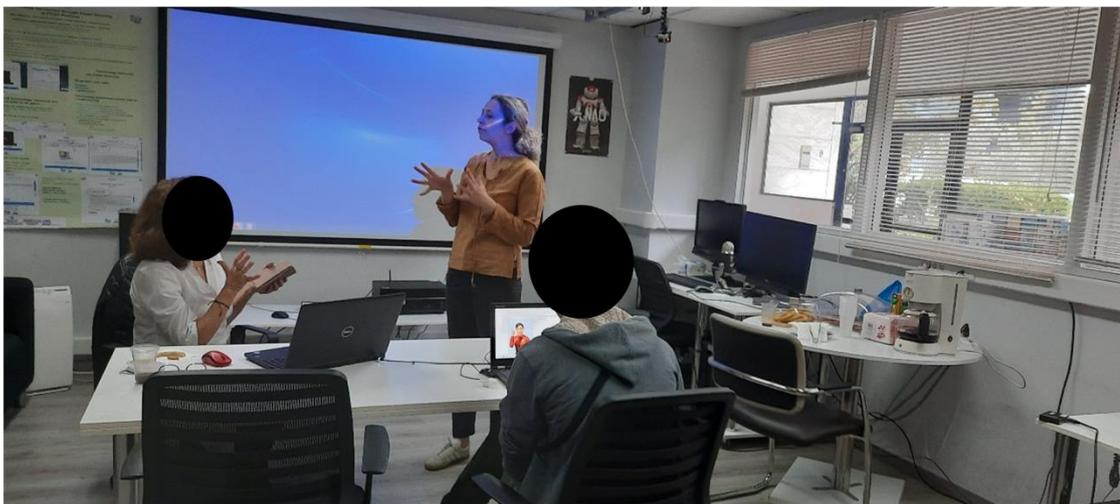


Figure 3: Facilitator providing explanations to deaf evaluator before the start of the app evaluation.

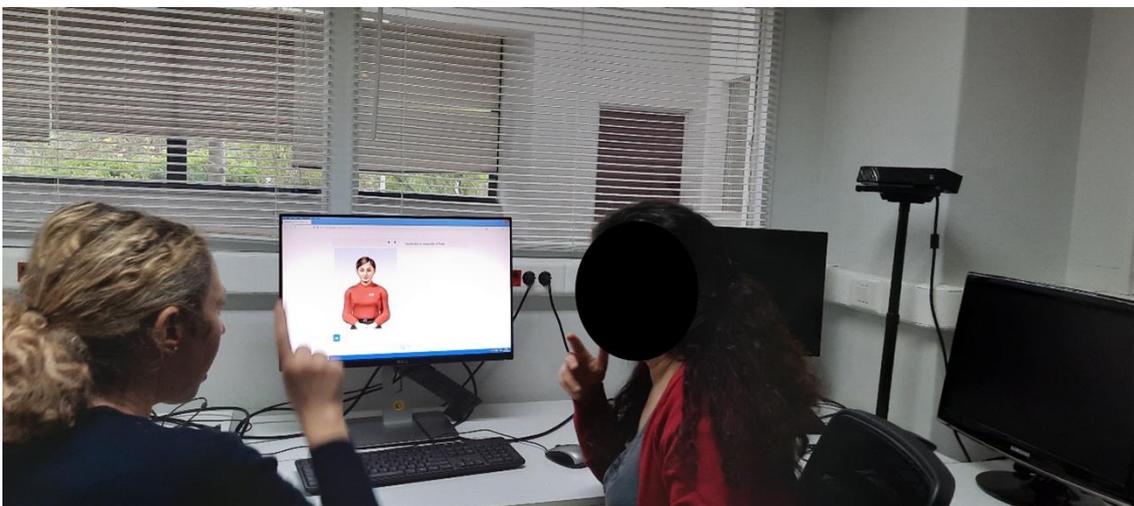


Figure 4: Facilitator with deaf evaluator during the avatar evaluation.



Figure 5: Instatiation from the discussion panel.

2 FOCUS GROUP DISCUSSION

2.1 App

2.1.1 General feedback

-Was the response generally positive, or negative?

In general, all participants agreed that the app is very useful and helpful for Deaf and hard of hearing persons since it will help Deaf people to have equal access to information and communication.

-What guidance did the facilitator give to the participants?

The facilitator clarified that the viewed screens depict the design of the app, which is not yet active and encouraged the focus group to navigate through all pages, make all allowed selections regarding language, avatar, profile and settings and spend as much time as needed to get familiar with the app and provide their comments regarding all aspects of its design.

-What did participants use to complete the questionnaire (mobile phone, laptop computer, etc)?

Participants used their mobile phones to navigate, but they found helpful also to have an overview of the app projected via the facilitator's laptop while they navigated, so that they could ask various questions on the app design, whenever they needed to do so.

-Which things did participants like, which did they not like?

Almost all participants reported some difficulty in interacting with the app. One major sources of difficulty was dealing with the settings of the application -as one evaluator mentioned "*it took a lot of effort to understand how the app works*". Almost all the participants suggested that a user guide should be available to let users consult it whenever needed.

All participants agreed that the design of the application is rather confusing since all of them faced some kind of difficulty while making the selection of input/output modes and languages. This coincides with remarks from the group of hearing evaluators, sine one participant stated that "*the application requires multiple back and forth clicks which is very time-consuming and confusing*".

Another participant suggested that a list of language names instead of flags would be more convenient for users.

Most participants found very puzzling the fact that after the selection of the input or output language the user should go back by clicking on the arrow. They suggested adding an "OK" button instead at the bottom of the screen in order for the user to move automatically to the next page or step.

One participant recommended having users' communication preferences saved at PROFILE.

One participant suggested that it would be nice to have the input and output functions as well as the input and output languages presented in a single page similar to the Google translate service by utilizing the screen horizontally instead of vertically. At the end of the evaluation discussion on the EASIER app, most of the participants

agreed in preferring to have an application that would look very much like Google Translate.

Concerning the avatar as embedded in the app, evaluators liked the options offered for customization, but they all agreed that good signing is the most important thing.

- Introduce major themes that came up (can add more than 3 Theme sections)

2.1.2 Theme 1: A simpler way interface is needed

The use of the arrow to move back and forth among selection pages for initial preferences should be replaced by an “OK” indicator, which would allow to move to the next step in a more intuitive way.

2.1.3 Theme 2: Incorporation of a manual

The evaluators pointed out that a user manual, which they could consult while using the app, would be a much preferred addition.

2.1.4 Theme 3: Horizontal use of the mobile screen

It was noticed that users would also prefer a horizontal use of the mobile screen. Since they are strongly influenced by the boxes providing input and output text in google translate, they would like to have a similar screen arrangement in the EASIER app as well. This was especially preferable in the case of translation from one SL to another.

2.1.5 Recommendations

- Did participants mention any specific recommendations for improvement?

Similar to the hearing group, deaf evaluators had no further recommendations for improvement other than the themes referred to above. They all agreed on the significance of the app for deaf-hearing communication and declared their interest to participate in other rounds of evaluation in the future, recognizing the importance of a tool like the app which is being developed in the framework of EASIER.

They also mentioned how pleased they were, that the EASIER team has asked for deaf persons' opinions, which they found very important, given that the app is planned to serve deaf users.

2.2 Avatar

2.2.1 General feedback

- Was the response generally positive, or negative?

All Deaf participants expressed a positive attitude towards the EASIER signing avatar. Also, they were impressed with the significant progress that the project team members have made in signing avatar technology, while two participants mentioned how good they found to have invited deaf GSL signers to validate the avatar's performance.

- What guidance did the facilitator give to the participants?

The facilitator made clear to evaluators that at the current stage they should concentrate in the hand activity of the avatar, since this is the part of the technology which is evaluated. But also encouraged them to express any suggestion that would provide feedback to the research work currently in progress.

-What did participants use to complete the questionnaire (mobile phone, laptop computer, etc)?

Just like the hearing evaluators' group, deaf evaluators used the laptop and desktop equipment of ATHENA to complete the questionnaire.

-Which things did participants like, which did they not like?

Evaluators agreed that the EASIER avatar is a high-quality avatar with a good signing performance. All informants stated that the vast majority of the signs produced by Paula are accurate and easy to understand. However, participants were also involved in discussing GSL linguistic issues such as right sign formation in the cases where they disagreed as to how a specific sign is formed. In these cases, they treated the avatar like a human and commented on the relevant GSL sign produced by the avatar providing recommendations and corrections that should be taken into account, since the current performance of these signs made them difficult to comprehend. Such signs included SORRY and SIGN, for which recommendations were provided as to handshape and place of articulation corrections.

Furthermore, participants noticed the cases of occlusion, just like the hearing evaluator had done, especially in the case of signs made with a 5-handshape or L-handshape when the thumb touches the chest, as for example in the GSL sign WAIT, where the handshape is not clearly visible. For such cases, they suggested a slightly rotated position of the avatar.

Three of the participants reported that the transition between the signs in signed phrases was a little bit abrupt and they suggested a "*relaxation of the hands among these transition movements*".

It is worth mentioning that the focus of the participants' criticism was not so much on the manual articulators but on the lack of the non-manual component, although they were informed that non-manuals are not part of this evaluation cycle. According to one of the participants the avatar was described as "*robot-like, emotionless and unnatural*" because of the absence of facial expressions, head and torso movement. All participants agreed with this remark.

In this line, the evaluators noted that there are some utterances which they were not able to comprehend because of the absence of facial expressions and head/body movements. As an example, they used the phrase NOT-UNDERSTAND, opening a discussion about how non-manuals should be present to make the phrase comprehensible. Thus, evaluators' suggestion for improvement of the avatar performance strongly focuses on the addition of non-manual markers since these features are important for the comprehensibility of sign languages utterances.

In the same line, participants noted the significance of mouthing and mouth gestures which are mostly used to disambiguate the meaning of a sign with the same hand configuration, using as an indicative example the sign SERVICE having two different meanings: a) service and b) employee, which are distinguished by mouthing. Two of the evaluators reported that they could not understand the exact meaning of the sign because of the absence of mouthing.

Parallel to non-manual markers, facial expressions for emotion (i.e., happy, angry, sad) has also been pointed out as an important element to be added. As one participant said "*I would like to see a smile when the avatar signs THANK YOU or GOODBYE*".

Finally, one evaluator suggested that it would be nice to make use of darker color for the avatar's clothing and background which make it suitable for deaf persons with Usher Syndrome.

- Introduce major themes that came up (can add more than 3 Theme sections)

2.2.2 Theme 1: Importance of incorporation of non-manuals in synthetic signing

As the deaf GSL evaluators group has pointed out, non-manual elements of signing cannot be separated from any SL representation that would be legible or acceptable by the signers' community. Thus, incorporation of non-manuals on both articulation and prosody levels is mandatory.

2.2.3 Theme 2: Occlusion is an obstacle for comprehension

Cases of occlusion which create ambiguity as to which handshape has been used in a specific sign formation form a critical source of lack of legibility. One suggestion is the slight rotation of the hand under specific articulation conditions, so that the handshape becomes visible and, thus, the utterance is disambiguated.

2.2.4 Theme 3: Darker avatar clothes and background to make it accessible to Usher Syndrome patients.

The remark for darker avatar clothes and background to make it accessible also to Usher Syndrome patients is a rather good point towards raising accessibility of the avatar by a wider audience.

2.2.5 Theme 4: Affect depiction is most welcome.

An evaluator has pointed out the preference for affect depiction with utterances where expression of affect is expected in real world to increase the avatar acceptance and naturalness.

2.2.6 Recommendations

- Did participants mention any specific recommendations for improvement?

Evaluators were satisfied with the evaluation process. They found the questionnaire an interesting way to watch and evaluate the avatar performance, as well as to commend on the various aspects of its articulation.

3 OTHER FEEDBACK

3.1 Feedback from facilitator (optional)

-How did the procedure go?

The procedure went very smoothly. All evaluators agreed that starting with the avatar evaluation and going next to the app evaluation has been a good course of processing. Similar to the hearing evaluators, they found this to be a good way to help them understand why we are developing synthetic signing avatar technology, while they all got excited with the perspective of a MT application. By the end of the session, also the deaf evaluators group expressed their interest to participate in the next rounds of the app and avatar evaluation whenever this will take place noticing that they appreciate a lot the organization of the evaluation and the participation of deaf people in the evaluation procedure.

What things could be changed for the final evaluation cycle?

It has been made clear by the discussion among evaluators and the relevant recommendations for avatar articulation corrections, that we should take care so that tokens which may raise theoretical linguistic discussions as to the way a sign is formed among different signers, are not included in the evaluation tasks. We have noticed that this may disorient evaluators who then start discussing how is a sign better/correct articulated and this is directly reflected to their judgment about the avatar performance, treating the avatar the way they would treat a human who does not use the right “words”.

Also a very careful preparation of the presentation of the tasks from the part of the moderators is absolutely necessary, to eliminate the option for disorientation of the evaluators group from the evaluation focus.



GREEK SIGN LANGUAGE HEARING GROUP

5. **NAME OF PERSON WRITING REPORT: ELENI EFTHIMIOU**
6. **DATE OF EVALUATION: 08 OCTOBER 2022**
7. **SIGN LANGUAGE: GREEK SIGN LANGUAGE (GSL)**
8. **GROUP: HEARING GSL INTERPRETERS AND GSL EXPERTS**

1 METHOD AND PARTICIPANTS

1.1 Facilitator

- Name of facilitator: Eleni Efthimiou
- Brief description of facilitator
 - internal to organization
 - hearing
 - language background: GSL researcher
 - professional background: Head of Sign Language Technologies Team at ATHENA/ILSP

1.2 Participants

- How many participants?
Five (5) experts (+1 in the pilot) participated in the evaluation of both the EASIER avatar and the app design.
- How were they recruited?
Participants were recruited by invitation among collaborators in other SL projects of the Institute.
- Describe the group with respect to age, gender, background/profile
The group was composed by three (3) experts working as GSL interpreters, one (1) GSL interpreter in probe period and one (1) university professor in Special Education with expertise in GSL as second language (L2). The expert participating in the pilot is a researcher with expertise in SL technologies, more precisely in GSL video recognition. The age of the group participants ranged between 25 and 55 years of age, with 2 participants in the 40-50 group. The group included five (5) female (including the pilot participant) and one male individual.
- Was there anyone else present in the session (other than participants and facilitator)?

In the session both hearing and deaf group facilitator were present. However, the deaf group facilitator did not take active part in the procedure.

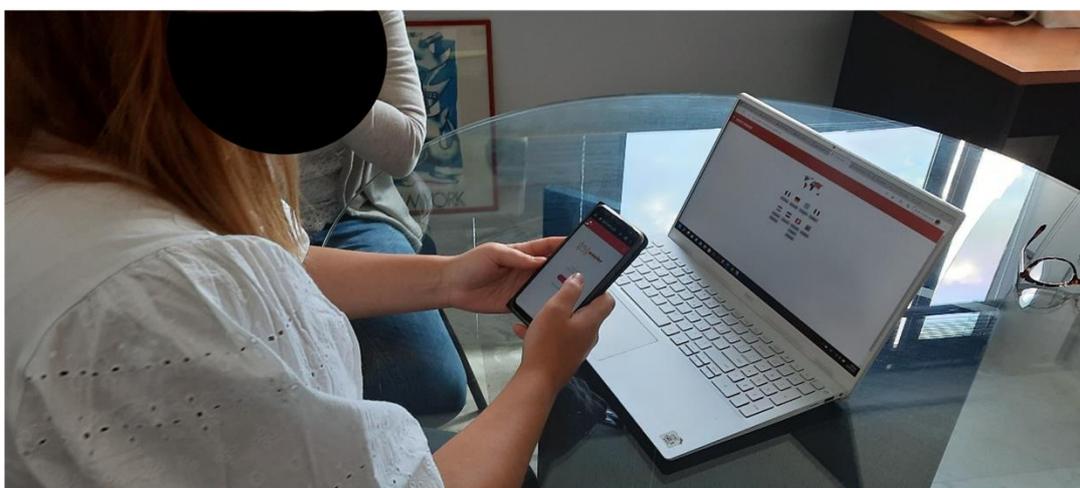


Figure 1: Inspection of the EASIER app during the pilot with a hearing GSL expert.

The two members of the technical staff controlling the cameras entered the session area once or twice when it was necessary to change the camera batteries or check camera condition, without disturbing the session.

1.3 Procedure

-Which components were tested?

Two components were tested: (i) the EASIER avatar, and (ii) the EASIER app design.

-How was the session organised?

First, participants were welcomed to the institute. Then we explained the background of the project. Then participants signed the related consent forms. Then we proceeded first with evaluation of the EASIER avatar and next with the EASIER app. The session was concluded with discussion among the evaluators and the facilitator on the user experience, the legibility of the avatar, the strong and weak points of the implementation so far and evaluators' suggestions as to what should be changed or improved during the next period in respect to the two evaluated components.

-What was the order of components?

Evaluators first viewed the EASIER avatar and then the app mock-ups. This had proved to be evaluation methodologically preferable, since they first got familiar with the SL representation engine of the project and its current state of development, so that they then could better follow the incorporation of the avatar in the app design.

-How long did each individual component take?

Introductory project presentation and consent form signing lasted about 15 minutes. The evaluation of each component had a duration of approximately 30 minutes, while another 30 minutes lasted the overall discussion following both components evaluation. This made a total of 90 minutes of evaluation time. A 10 minutes break was made between evaluation of the avatar and the app. Another break of equal duration preceded the discussion session.

-What was the total duration of evaluation?

The overall duration of the evaluation was approximately 125 minutes.

1.4 Technical setup

-Was the evaluation online or in person?

Four participants were present in the premises of ILSP/ATHENA where the evaluation was conducted. One participant was connected online and participated in all phases of the procedure via Skype teleconferencing.

-What equipment was used?

- For interacting with the components there were used the evaluators' smartphones, two institution laptops and two institution desktop devices.
- For recording participant discussion, three GoPro cameras were used mounted on the ceiling of the evaluation room, as well as one HD camera on a tripod to capture a global view of the room. All cameras were High Definition, recording in 1920*1080 pixels and 50fps (frames per second).
- Other equipment involved a projector, via which the remote participant participated in all phases of the evaluation and also the facilitator projected the EASIER app.

-What did the setup look like?

The camera and discussion panel setup is depicted in the following figures.



Figure 2: Evaluation room with GoPro setup. L indicates left side camera, C indicates central camera, R indicates right side camera.



Figure 3: Facilitator and participants during app inspection via evaluators' smartphones. Parallel projection of the app mock-ups facilitated evaluators' approach to the app design.



Figure 4: Deaf and hearing facilitators with the evaluators panel including the remote participant.

2 FOCUS GROUP DISCUSSION

2.1 App

2.1.1 General feedback

-Was the response generally positive, or negative?

The response of the evaluators was generally positive.

-What guidance did the facilitator give to the participants?

The facilitator clarified that the viewed screens depict the design of the app, which is not yet active and encouraged the focus group to navigate through all pages, make all allowed selections regarding language, avatar, profile and settings and spend as much time as needed to get familiar with the app and provide their comments regarding all aspects of its design.

-What did participants use to navigate (mobile phone, laptop computer, etc)?

Participants used their mobile phones to navigate, but they liked the idea of having also an overview of the app projected via the facilitator's laptop while they navigated (Figure 3), so that they could ask various questions whenever they needed to do so.

-Which things did participants like, which did they not like?

Participants liked the design of the app in general. They stated that after spending a few minutes with the app they had no difficulty in understanding how it works and that it was clear to them.

However, they would prefer to have fewer transitions back and forth while choosing input/output language.

Another comment was that it would be preferable not to be obliged to scroll down the app page in order to find function buttons like the "start" button. But this may be due to the fact that even from their mobile phones they were viewing a webpage and not the real app.

-Introduce major themes that came up (can add more than 3 Theme sections)

2.1.2 Theme 1: Visible functionality keys

There was a strong preference/recommendation among evaluators for fixing the screen so that all function buttons, especially referring to the "START" button, are visible without having to scroll the screen up and down.

2.1.3 Theme 2: Always visible settings

Evaluators expressed a preference for having access to settings also after they have made their initial choice, possibly through a visible settings button.

2.1.4 Recommendations

-Did participants mention any specific recommendations for improvement?

Evaluators had no further recommendations for improvement other than the two themes referred to above. They strongly mentioned the significance of the app for deaf-hearing communication and declared their interest to participate in other rounds of evaluation in the future, recognizing the importance of a tool like the app which is being developed in the framework of EASIER.

2.2 Avatar

2.2.1 General feedback

-Was the response generally positive, or negative?

The response of the evaluators was generally enthusiastic. They all mentioned the significant improvement of the avatar performance in respect to technologies they had viewed in the past. In parallel, they did mention the need to add facial features and upper body motion to make the signed message more clear.

-What guidance did the facilitator give to the participants?

The facilitator made clear to evaluators that at the current stage they should concentrate in the hand activity of the avatar, since this is the part of the technology which is evaluated. But also encouraged them to express any suggestion that would provide feedback to the research work currently in progress.

-What did participants use to complete the questionnaire (mobile phone, laptop computer, etc)?

The participants used the laptop and desktop equipment of ATHENA to complete the questionnaire.

-Which things did participants like, which did they not like?

The participants liked a lot the performance of the avatar in general. They had some disagreement/discussion among them about how specific signs are formed in GSL, which ended with recommendations about the content but not the quality of the avatar signing.

-Introduce major themes that came up (can add more than 3 Theme sections)

2.2.2 Theme 1: Incorporation of prosody

All evaluators expressed their interest to see the next steps of the avatar performance, when intra-sign facial features and sentence-level prosody will be incorporated. This is generally agreed to be a necessary step towards full legibility of the avatar signed productions.

2.2.3 Theme 2: Handling of occlusion

All evaluators noticed that cases where 5-handshape and L-handshape form signs in front of the signer or in direct contact with the signer's body at mid-chest level are equally difficult to distinguish both in human and avatar signing. One solution could be to rotate the hand a bit so that the handshape becomes visible.

2.2.4 Theme 3: Smoothing of transitions within signed phrases

One reviewer mentioned that transitions from one sign to the next within signed phrases should be smoothed so that the overall performance becomes less "robotic".

2.2.5 Recommendations

-Did participants mention any specific recommendations for improvement?

Evaluators were satisfied with the evaluation process. They found the questionnaire an interesting way to watch the avatar performance from different perspectives and commend on the various aspects of its articulation.

Regarding the avatar component, they generally agreed on the remarkable progress that has been made in signing representation and noticed that if the themes mentioned above can be incorporated in the technology under development, they could see signing avatar incorporation to more demanding tasks like the teaching of SL. They also mentioned

that the translation app would be a very helpful tool even as a look-up tool for isolated sign translation purposes.



3 OTHER FEEDBACK

3.1 Feedback from facilitator (optional)

-How did the procedure go?

The procedure went really smoothly. All evaluatots expressed their preference for starting with the avatar evaluation and going next to the app evaluation. They found this to be a natural way to help them understand why we are developing signing avatar technology, while they all got excited with the pespective of a MT application. It is characteristic that by the end of the session, they all expressed their interest to participate in the next rounds of the app and avatar evaluation wherever this will take place.

-What things could be changed for the final evaluation cycle?

We should take care so that tokens which may raise theoretical linguistic discussions as to the way a sign is formed among different signer groups, are not included in the evaluation tasks. This may disorient evaluators who then start discussing how is a sign better/correct articulated and this is directly reflected to their judgment about the avatar performance, treating the avatar the way they would treat a human who does not uses the rght "words".

Also a very careful preparation of the presentation of the tasks from the part of the moderators is absolutely necessary, to eliminate the option for disorientation of the evaluators group from the evaluation focus.

ITALIAN SIGN LANGUAGE DEAF GROUP

**7.9 NAME OF PERSON WRITING REPORT: DON KULATHUNGA
(TRANSLATED BY DAVY VAN LANDUYT)**

7.10 DATE OF EVALUATION:

7.11 SIGN LANGUAGE: LIS

7.12 GROUP (DEAF/HEARING): DEAF



1 METHOD AND PARTICIPANTS

1.1 Facilitator

Don Kulathunga

1.2 Participants

6 participants, 3 women and 3 non-binary, between age 23 and 30. Their profiles are psychology student, unemployed traveller, working in a bar, as a gamer, student/actor and actor.

1.3 Procedure

There was a 30-minute explanation, then 40 minutes of testing and afterwards 70-80 minutes of discussion.

1.4 Technical setup



2 FOCUS GROUP DISCUSSION

2.1 App

The starting page looks simple enough, but there was some confusion about the word Start in the upper left corner, and the Start button.

Then after clicking Start, it was not clear which choice it was providing, is it the language the user is using or is it the output language?

By proceeding with language selection, you need to keep clicking on Back, which is not an intuitive linear progression. The Back button also brings you completely back to the start, it should bring you back to the previous screen within this linear process, so that users can go one step back or forwards.

The participants also did not notice any “LIS” when selecting camera, instead it shows “Italian” which means their sign language is not represented in the app.

Participants suggested to have a video manual on how this app works, to explain the progress, so that people with difficulties reading can use the app as well, since it’s not intuitive. Now you have to make guesses yourself and try to figure out how it works.

The map image in the language selection screen was confusing. What does the red areas in the image mean? Later in the settings, there are also two identical images, do those mean anything? Are they related to each other? It was confusing to the participants. What is the difference? The images should be different to separate them.

Participants also wondered about the presence of Asian text characters in the language selection button - why are there Asian characters, as this is an EU project and those languages are not included? Does it mean that Asian languages are also an option to translate to?

The participants saw the recurrence of the color red over and over in the images through the translation process and are trying to find out what it means. They did not understand the meaning of the image of the red hand holding the phone with the arrow to the question mark. They felt like if the color red had any specific meaning. The format of the art is also quite small.

The avatar settings: they did not understand the difference between the Contrast and the Background settings. What is the difference between those two? Clicking on those options doesn’t give any examples, they cannot see visually what the change is. The avatar image stays the same, so clicking around here feels pointless.

The images of the suit and the t-shirt were also not liked - it is not gender neutral. Men have a suit, what about women? It should be gender neutral. They also commented that the t-shirt option is not for deaf blind people, they should have long sleeves. The color options for the shirts should also be more simple, just a white or a black option.

The avatar is also white, there is no skin colour diversity. If BIPOC participants see a white interpreter, they don’t feel represented and they feel excluded. Don’t forget that there are BIPOC in Europe as well!

Participants also wondered if the avatar needed a logo on her chest, maybe it could be impeding the visibility for some. A good option would be to have the logo in a upper corner.

The info page also does not mention deafblind individuals along with “deaf” and “hearing”, so are they not included?

When you have enabled the dark mode, the screen flashes briefly white during navigation between the different screens. This is hard on the eyes during navigating in dark mode.

There are also issues with black and grey colors in the dark mode, for example the black stripe of the German flag disappears in the dark mode or the text under the EASIER logo. Settings along the dark mode and text size should also include boldness of the text, that you can make it bolder. Participants compared it to subtitle settings on Netflix where you can choose text color (e.g. yellow or white), size, boldness. Don't force one choice to the users.

The gender pictograms for men, women and "X": the X was not accepted by the participants, they felt that this was stigmatising. They preferred "other" or a different image. For example intersex people did not fit in any of the options. They felt this app represented the privileged persons: white cis persons.

Participants felt lost in the navigation, with all the settings, back and forth etc. After making a translation, it was not clear what the buttons Copy, Save meant. They felt the need for explanation in the app to make the functionalities clear to the users.

There are also issues with the alignment of some design elements. For example in the Settings screen, the icons for input and output are not aligned properly.

There was a suggestion for the background color of the buttons to change when you hover over it with the mouse, so it is a visual support.

The world map image: why is it not only Europe? It feels weird to see the world as it's an European project.

There are questions regarding the video data after the recording. What happens with it? Is it stored or removed? Is it stored on my phone? What does the save button afterwards do?

Facilitator noted that participants had difficulties navigating, it was not clear for them.

Participants noted that the format was very small, it should be bigger.

Participants suggested a FAQ section, with contact. Who can they contact with questions? Where can they go if they forgot their password?

One participant asked if there are any requirements for the video recording? How does the user know when the video is of good quality for the translation? E.g. wearing fingerless gloves because it is cold, having half the face hidden behind a shawl, wearing big earrings, etc. Will the app understand the signing, or are there certain requirements that need to be met so that the app will understand me well? There is currently no indication.

Participants commented that they needed more clear explanation about (the differences between) the buttons. The navigation was not intuitive. More information is needed.

3 OTHER FEEDBACK

3.1 Feedback from facilitator (optional)

- How did the procedure go?
- What things could be changed for the final evaluation?





ITALIAN SIGN LANGUAGE HEARING GROUP

- **NAME OF PERSON WRITING REPORT: LUCA MARRA**
- **DATE OF EVALUATION: OCTOBER 28, 2022**
- **SIGN LANGUAGE: ITALIAN SIGN LANGUAGE**
- **GROUP (DEAF/HEARING): HEARING**



6 1 METHOD AND PARTICIPANTS

1.1 FACILITATOR

- Luca Marra
 - SWISS TXT
 - hearing
 - Lis/Italian
 - Interpreter, researcher

1.2 PARTICIPANTS

- 5 participants
- The group was made by 4 female and 1 male, everyone between 23 and 40 years old.
- We tested the app both on mobiles and laptops, since usually people like to use apps also on the desktop when it is possible.
- How was the session organised?
- The session was divided in this way:
 - Easier presentation
 - Description of the evaluation parameters
 - Evaluation on the mobile
 - Evaluation on the desktop/laptop
 - Overall debriefing
 - Explanation and info about the vouchers
- In total the session lasted two hours and half:
 - 30 minutes presentation and parameters
 - 60 minutes testing
 - 60 minutes discussion

1.3 TECHNICAL SETUP

- The session has been made online via ZOOM
- Each participant was connected through desktop/laptop and each one of them has a mobile available for testing the app



- Each participant created a document during the evaluation in which they write down the things they observed during the test.
- Everyone was using his own desktop/laptop with his own camera built in



2 FOCUS GROUP DISCUSSION

2.1 APP

2.1.1 General feedback

- The group was generally not happy with the app. In general, they complained about 4 aspects:
 - The UI interface was hard to understand
 - It was not easy to interact with the UI interface
 - There are a lot of translation problems
 - General problems about the design/layout/icons

- The facilitator asked one participant to start with observations, starting with his or her paper, and then asked the other participants to speak on a point made when they also had observations on that point, this was to avoid repetition and redundancy.

Once the first participant's remarks were finished, the second participant moved on, but of course in this way the topics already discussed were not dealt with again.

In this way, the observations were shared in a fluid manner, at the same time containing the different points of view of all participants.

These feedbacks will be now described in the section below.

2.1.2 Theme 1: INTERACTION, INTUITIVITY, STYLE AND INCLUSION PROBLEMS

- The map confuses the users, since choosing a language the enlightened country doesn't change





- The icon chosen for the non-binary gender can be seen as a discrimination, since usually the “x” is perceived has a negative/bad thing.

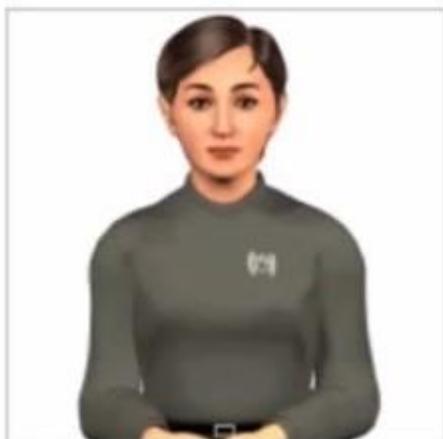
Genere



- The avatar doesn't have the possibility to be chosen with different skin colors, this can be perceived as an ethnicity discrimination.



- Different background and cloth colors are not working.



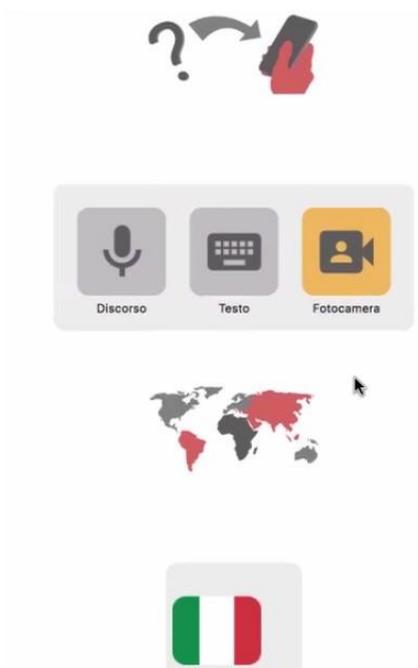
Sfondo	Abbigliamento	Genere
<input type="radio"/> 	<input type="radio"/> 	<input type="radio"/> 
<input checked="" type="radio"/> 	<input type="radio"/> 	<input type="radio"/> 

- Switzerland has 3 written languages, from the icon is not clear which one is chosen.

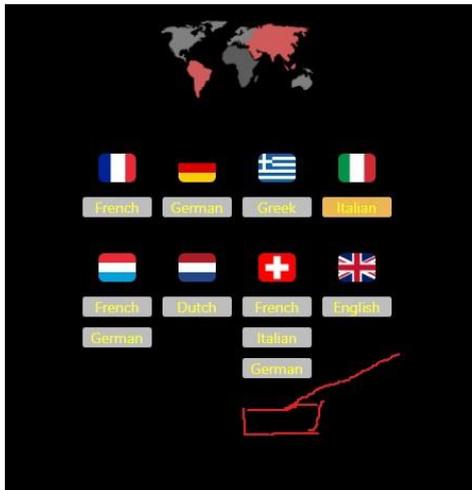




- When I choose a language for an input (for example) the system sets the same language for all the "input/output", that is not ok.



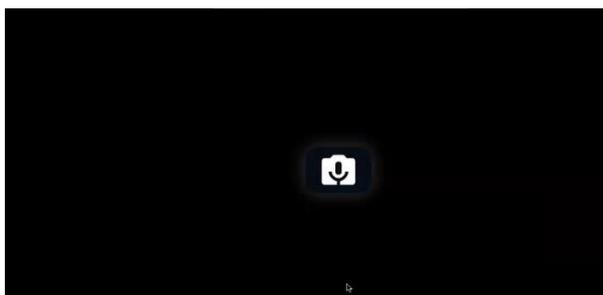
- Every user was not happy with the fact that after every choose they made they had to press on the "back" button, it would be better having an "ok" button or something like that after every choose for continuing and avoiding the "back" button.

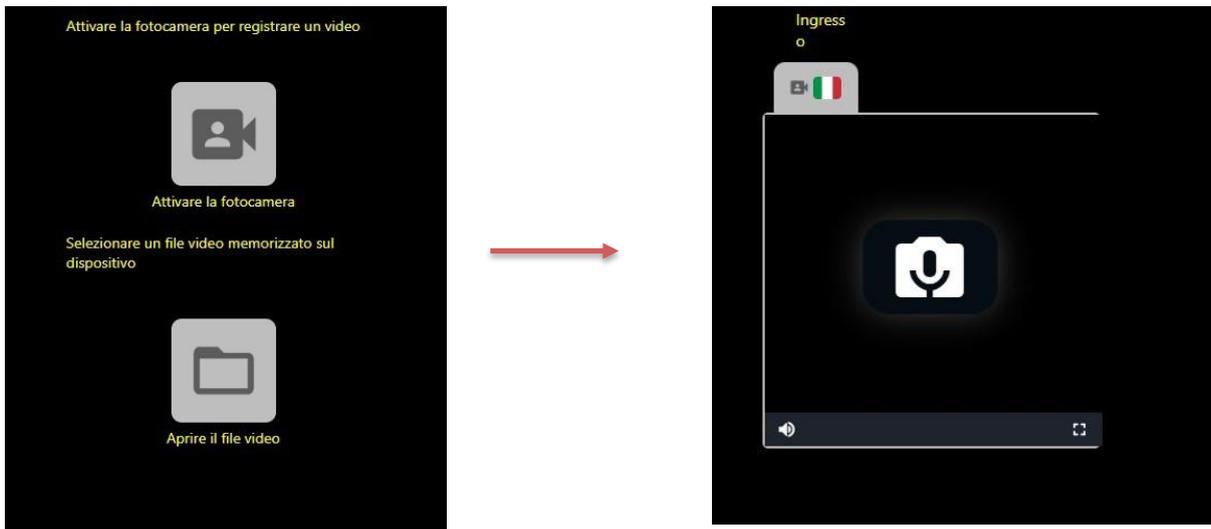


- If you press “back button” from the browser and not the “back button” inside the app, the screen goes white, and you have to start again, please fix this.

- If you go on “full screen mode” the app crashes:

- When you have to activate the camera, the systems asks you to do that two times, one time is enough, please fix this.



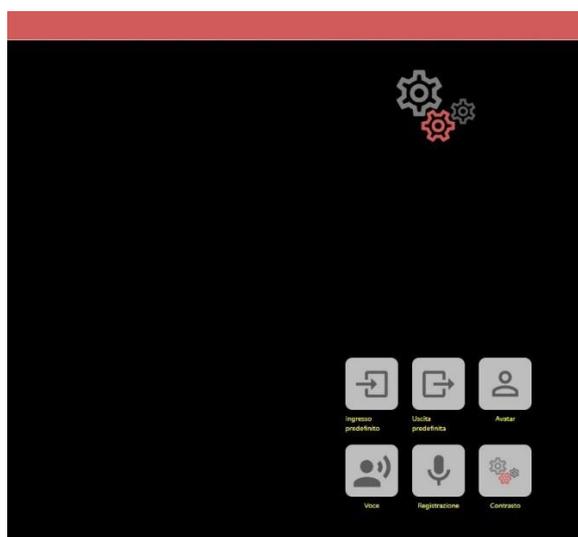
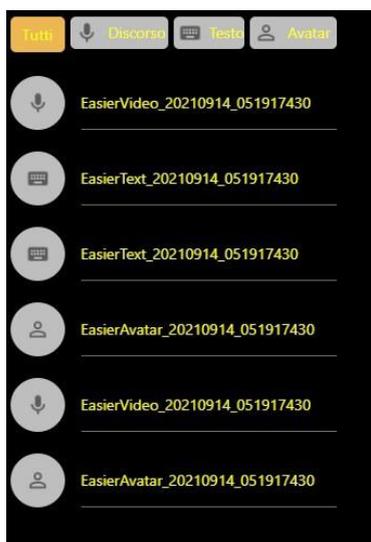


- The same icon is used for 2 different functions: “settings” and “contrasto”, that is not intuitive.
- The menu colors don't match with the whole system color identity. Why using red for

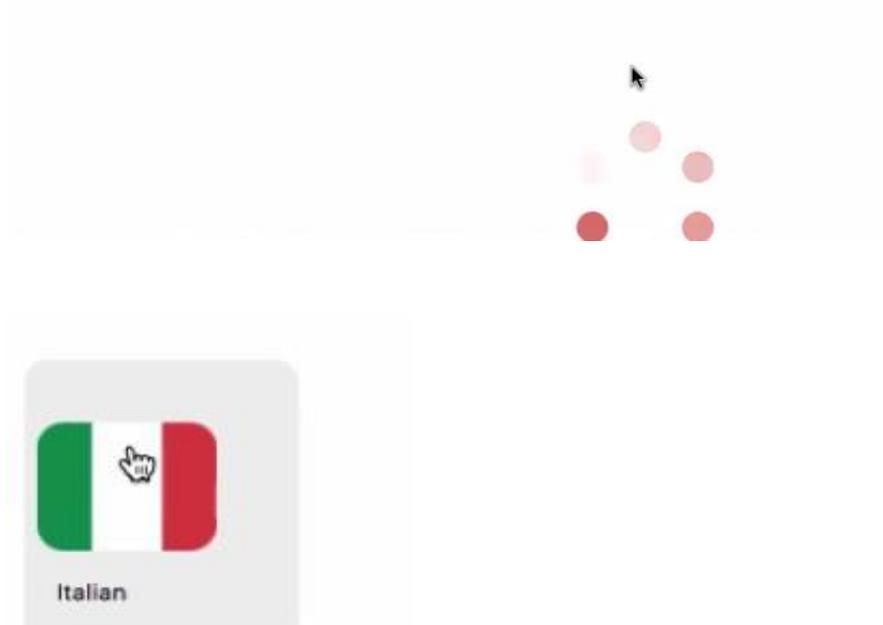


the entire menu and then some text is in yellow and not in red? Or in a color that is a different red nuance, to maintain the colour identity of the project?





Accesso a

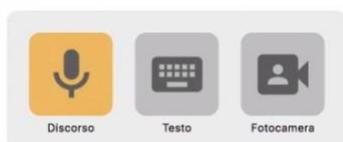


- The order of the icons is semantically wrong, it would better having “Voice, Avatar and recording” on the same line.





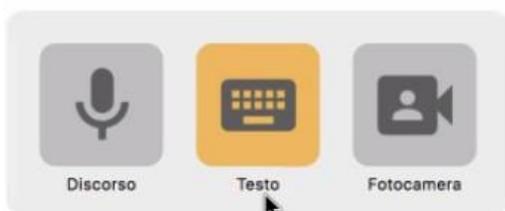
- It would be better if in one screen you choose the modality and in the next page that appears you choose the language, otherwise is not intuitive:



- “impostazioni preferite” gave some problem to the users, what it the real function?



- The name of this line of icon is not semantically correct. Or you choose the “device” or you choose the “function”. So or you put “Speaking, writing, signing” or you put “Microphone, text, Camera”. But this mixing is not good.



- It would be great having in every page this horizontal menu bar, so you don’t have to go back to the home page to find these functions:



- The elements on this page are not clearly distributed. Too many text lines and positioning between the elements is not the same (the empty space measurement changes):



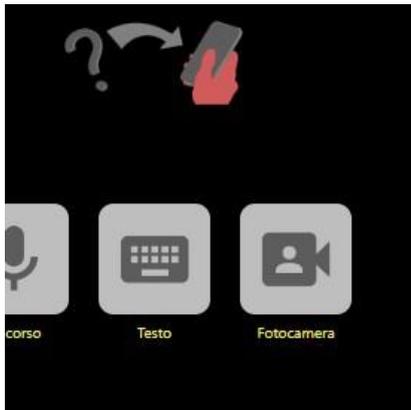
- The fact that the words are divided on two lines is not good for the eyes perception:



- The rectangles size don't fit the words, or viceversa. Please fix it:



- The icon of the smartphone can be inadequate if I am using the app from my laptop/site. Since everyone uses app sites, we should be clear from the beginning with the icon:

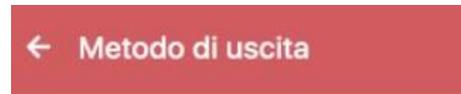
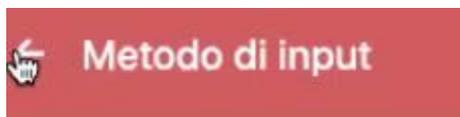


- The icon in general has an old style, 90's 2000's style. Since the project is so innovative, we should use a style that is innovative too:



2.1.3 Theme 2: TRANSLATION PROBLEMS

- The focus group noticed that sometimes there is no coherency between the different “pages” of the APP, things are called in a different way from one time to another. So for example if you call “input method” one page, you should name the other “output method” and not changing the syntax:



- The accent section has no meaning in other languages in this way (british/american), you could either delete that part, or you could adapt the accent on the specific need of that language:





- If I choose a language, also the languages should be translated into that language, some part of the app is still in English:



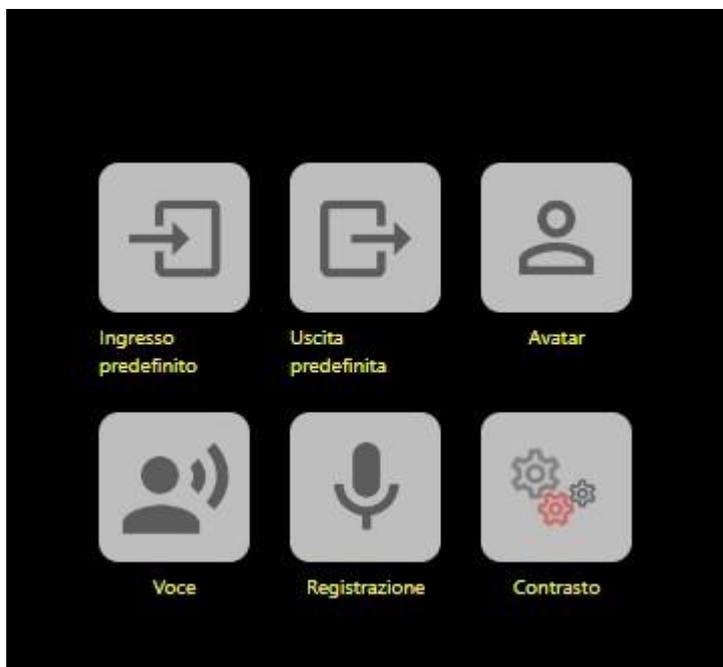
- If I change the language, also the alphabetical order of the words in that languages should change. As you can see here, the G for German is after the I for Italian, is it clear that it follows the alphabetical order of the German Language:



- Here the translation of the icons makes people think that “ingresso predefenito” and

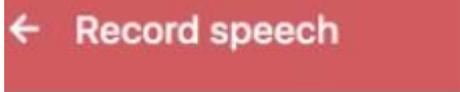


“uscita predefinita” are for peripherals devices (microphone, speakers etc), is not clear that they are referring to the language output:

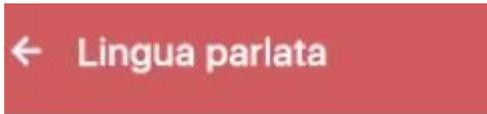


Impostazioni delle lingue



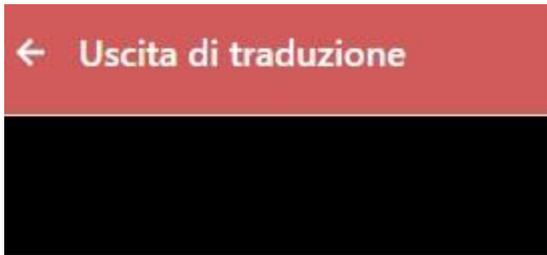


← Record speech



← Lingua parlata

- This sentence has no meaning in Italian, it's a calque:



← Uscita di traduzione

- This sentence It's weird in Italian, we don't use "touch" so it's better leaving just "Scrivere":



- Here there is another coherence issue, in a page you have some words, in the following page you have different words for the same functions:



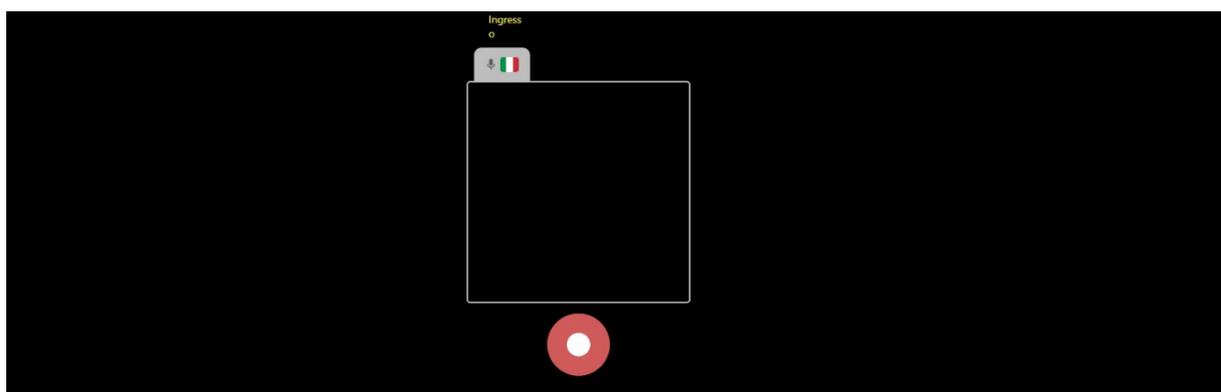
- Here there's another coherence conflict; you choose camera and then between the languages there's "Italian" and not "LIS":



- Here a translation issue, but also the icon is misleading. The translations says "is the text clear?" but the real question is "Are you sure/do you want to continue"? The icon with the red X, is not coherent a from an iconic point of view.

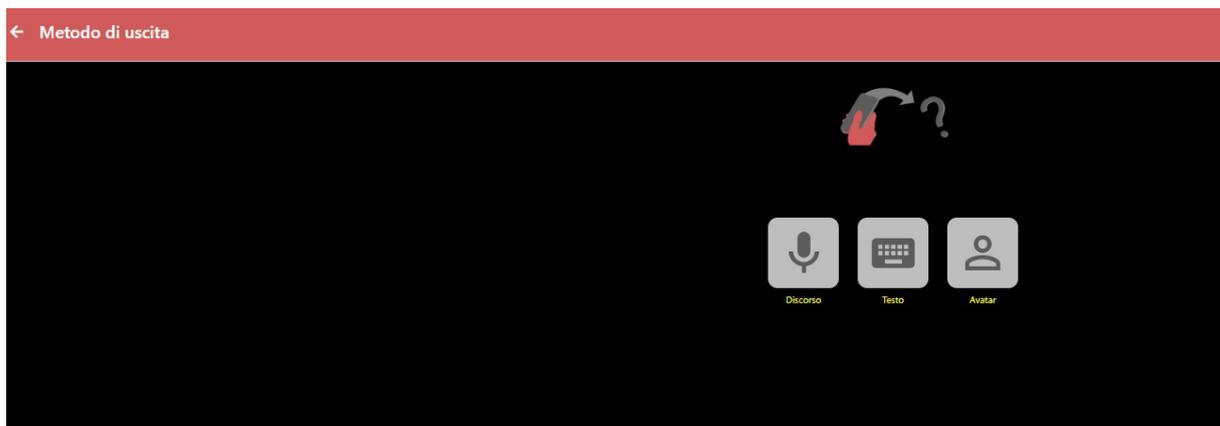


2.1.4 Theme 3: OTHER MOBILE/DESKTOP PROBLEMS



- From a desktop view, here there is too empty space:





- From the mobile view, here the menu with the languages is not visible entirely:

