D1.4 REPORT ON FINAL EVALUATION STUDY

Revision: v1.0

Work package	WP 1
Task	1.3
Due date	31/12/2023
Submission date	9/12/2023
Deliverable lead	EUD
Version	1.0
Authors	PICRON Frankie, VAN LANDUYT Davy, OMARDEEN Rehana (EUD), EFTHIMIOU Eleni, FOTINEA Evita, WOLFE Rosalee, GOULAS Theo, VASILAKI Kiki (ATHENA), MORYOSSEF Amit, MÜLLER Mathias, EBLING Sarah (UZH), TISMER Christian (NURO)
Reviewers	BONREPOS Claire (INT), BRAFFORT Annelies (CNRS)
Abstract	This Deliverable is the report which presents the results from the final phase of evaluation studies in the EASIER project. The end-user evaluations were carried out by WP1 partners, with both deaf and hearing evaluation groups. Components from the EASIER project (application, machine translation systems 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

Document Revision History

Version	Date	Description of change	List of contributor(s)
0.1	15.01.2021	Template	Martel
0.2	10.11.2023	First draft	EUD
0.3	23.11.2023	Additions from responsible partners for evaluation components	ATHENA, UZH, NURO
1.0	08.12.2023	Final version incorporating reviewers comments	EUD

DISCLAIMER

COPYRIGHT NOTICE

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.

Funded by the Horizon 2020 Framework Programme of the European Union

EXECUTIVE SUMMARY

This report presents the results of the final 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 Tasks 1.2 and 1.3 with the aim of ensuring that end-user feedback can be collected and integrated into the design of the EASIER system.

The final evaluation was carried out across 5 countries with 96 deaf and hearing participants from the project's target sign language communities. Three components of the EASIER system were evaluated: the app, the translation models and the avatar. The app and avatar were evaluated in facilitator-led groups; there participants interacted with these systems, performed a rating task and then engaged in a facilitator-led discussion group to provide in depth feedback and recommendations. The translation models were tested using an online rating paradigm which participants completed over the course of several weeks. Upon completion, participants were also invited to give more qualitative feedback on the translation models.

The app prototype received an average usability rating of 60.1/100, which can be classified as "okay", however scores ranged from 80s (good) to the 40s (poor) with hearing participants rating the app on average slightly better than deaf participants. Focus group discussions revealed specific issues with respect to app organisation and settings. While indicating several clear fronts for improvement, qualitative feedback on the app also demonstrates the large range of personal preference when it comes to mobile applications.

The avatar showed global improvements in acceptability and readability across all groups, a finding that ran through both quantitative and qualitative results. Participants were able to directly appreciate the improvements by judging new and old avatar versions side by side. Nevertheless, they stressed the need for improvement specifically with respect to movements of the face, mouth and body to increase both acceptability and legibility of Paula's signing.

As expected, the output of the machine translation models performed consistently worse than human translators in all languages and all translation directions. Nevertheless, in the sign to text direction, models showed some (limited) promise with ratings consistently above 0. While the scores are very low, these results demonstrate a basic capacity that will be built on in future work, whether by refining current approaches or by exploring new approaches. Combining qualitative and quantitative feedback also highlighted some technical and methodological issues that may have led to lower ratings and that can be refined in future work.

The evaluation has combined quantitative and qualitative methods to both establish benchmarks of user acceptance for the existing technology within EASIER as well as create a roadmap of how to continue working to improve these technologies, during the final months of the project and beyond.

Funded by the Horizon 2020
Framework Programme of the European Union

TABLE OF CONTENTS

Funded by the Horizon 2020 Framework Programme of the European Union

LIST OF FIGURES

Funded by the Horizon 2020 Framework Programme of the European Union

LIST OF TABLES

Funded by the Horizon 2020
Framework Programme of the European Union

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
IS	International Sign
LIS	Italian Sign Language
LSF	French Sign Language
MT	Machine translation
NGT	Sign Language of the Netherlands
SL	Sign Language

Funded by the Horizon 2020
Framework Programme of the European Union

1 INTRODUCTION

The final user evaluation study falls under Task 1.3, **End user evaluation studies**, of Work Package 1. The final user evaluation is the second of two user evaluations planned in the project lifespan, with the first one taking place in M21-M23 and resulting in D1.3 **Report on interim evaluation study** (M24). The aim of these evaluations is to collect end-user feedback from the project's target language communities on the components of the EASIER system, specifically the mobile application, the translation models and the sign language avatar.

Certain adjustments to the evaluation have been made from the original project proposal, as well as the plan outlined in Deliverable 1.2 **Report on performance metrics and user study preparations**. First, one sign/spoken language pair, NGT/Dutch, has been dropped from the evaluation because the responsible partner, Radboud University, has left the project, leaving us with no appropriate partner to carry out this evaluation. Second, the number of participants has been reduced from 50 to between 20-24 participants per language group (10-14 for app/avatar, 10 for MT). This decision has been made to allow us to collect more focused feedback for each component, incorporating both qualitative and quantitative perspectives and when necessary, as in the case of the MT component, recruiting participants with specific profiles.

In order to collect feedback on the different components under development from members of the different language communities, we developed an 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.

The app and avatar component evaluations closely followed the method of the interim evaluation study with facilitator led questionnaire and focus group discussions. These are described in Sections 2 (Method), 3 (App results) and 4 (Avatar results). The machine translation component was evaluated fully online by translation professionals, and is described in Section 5. Section 6 concludes.

Funded by the Horizon 2020
Framework Programme of the European Union

2 APP & AVATAR METHOD

2.1 INVESTIGATION METHOD

The app and the avatar components were evaluated in a facilitator-led group setting, following the method of D1.3. Here, evaluations consisted of a combination of qualitative and quantitative feedback. For both the app and the avatar, participants were first shown the current state of the technology and asked to complete a structured rating task. After this, facilitators led a group discussion to get more in depth qualitative feedback about the technology. This method allowed us to not only get global benchmarks for how the technology is viewed by users, but also collect invaluable feedback on how to best improve things to achieve maximum user acceptance.

The final evaluation phase of the app and avatar took place over the period of September-October 2023, and included in total ten groups from five sign language communities. For BSL, there was no avatar available so only the app was evaluated. The rest of this chapter describes the user recruiting process and the procedure of the app and avatar 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), Greek Sign Language (GSL), and French Sign Language (LSF). For each of the 5 communities, there were two separate evaluation groups, one with deaf and one with hearing participants, resulting in a total of 10 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 the case of GSL a hearing project member who is a CODA and a long-standing member of the signing community acted as facilitator for the deaf group. 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. All but two facilitators reprised their role from the interim evaluation; the two new facilitators that joined in this round both took part in the interim evaluation as participants.

2.1.1.2 Participant recruitment and profile

For each group, between 5 and 7 participants were recruited who use the target sign language. There was no specific professional or educational background required for participants; however for those evaluating the avatar, a high degree of fluency in the relevant sign language was a requirement.

Facilitators along with the local partner were responsible for recruiting evaluation participants and recruitment was carried out through personal and professional networks. Across all groups, some participants who took part in the interim evaluation were invited back for the final evaluation. This was done because these participants were able to more accurately judge the progress made from the previous round. Participants received financial compensation for their participation, in line with each local partner organisations' rates and guidelines.

The recruitment resulted in a total of 59 participants; 29 across all deaf groups and 30 across all hearing groups. One participant identified as deafblind. Participants were 11 men and 36

Funded by the Horizon 2020 Framework Programme of the European Union

women. Most groups across all languages included at least one participant who also took part in the interim evaluation.

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: Summary of participants and groups for facilitator-led evaluation

Language	Partner responsible	Components tested	Group	N° of participants	Setting
BSL	DCAL	app	deaf	**6**	online
			hearing	**5**	online
DGS	UHH	app, avatar	deaf	**6**	face-to-face
			hearing	**5**	face-to-face
DSGS	UZH	app, avatar	deaf	**6**	face-to-face
			hearing	**6**	online
GSL	ATHENA	app, avatar	deaf	**7**	online & face-to-face
			hearing	**6**	online & face-to-face
LSF	INT	app, avatar	deaf	**5**	face-to-face
			hearing	**7**	face-to-face

2.1.2 Procedure

Given that the procedure was almost identical to the interim evaluations and all facilitators had either previously facilitated or participated in the last round of evaluations, we eliminated the pilot stage. EUD created updated guidelines for the evaluations based on feedback from the interim evaluations and distributed these materials and met with facilitators to ensure everyone was well-informed before beginning the final evaluation round.

Given the scale of the evaluation and the number of partner institutions, each facilitator 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. Several

Funded by the Horizon 2020 Framework Programme of the European Union

partners conducted multiple small group evaluations (DSGS hearing, GSL deaf, GSL hearing) to optimise scheduling participants.

For those groups that were conducted online, participants used their own devices (either mobile phones or computers) to navigate the online app and app questionnaire as well as the avatar questionnaire. 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.

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, erased after use and not shared with any other consortium members.

In most groups, the facilitator and participants were the only ones present in the room during evaluation, but in some cases technical staff also assisted with video recording of sessions. However, for some groups, the facilitator for the other group was also present to take notes. Evaluation sessions with deaf groups were conducted in the local sign language, and sessions with hearing groups were conducted in the local spoken language.

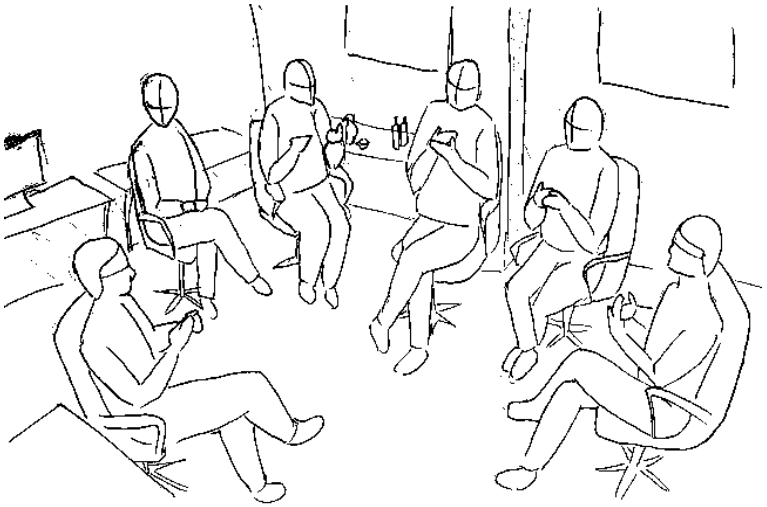


FIGURE 2.1: DGS setup for evaluations. Participants used desktop computers in the testing room and sat in a semicircle for discussion with a large screen behind the facilitator for showing materials.

Funded by the Horizon 2020
Framework Programme of the European Union

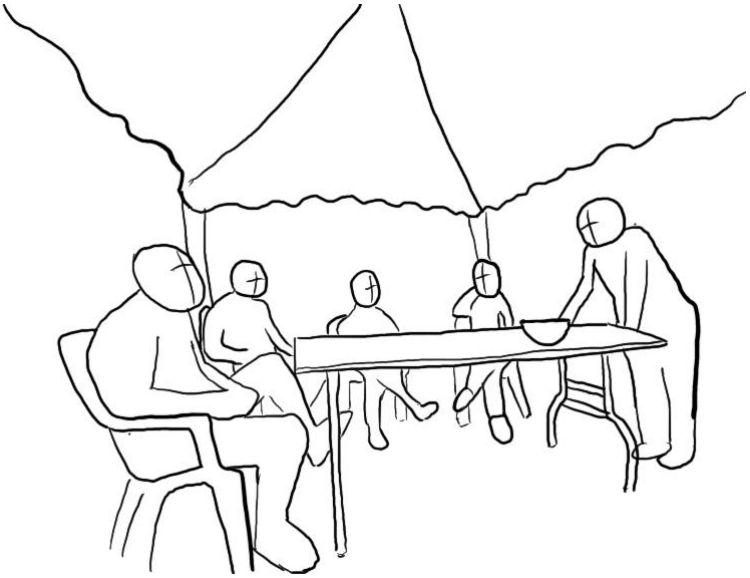


FIGURE 2.2: LSF setup for feedback discussions. Testing was conducted inside and discussions were conducted outside with refreshments

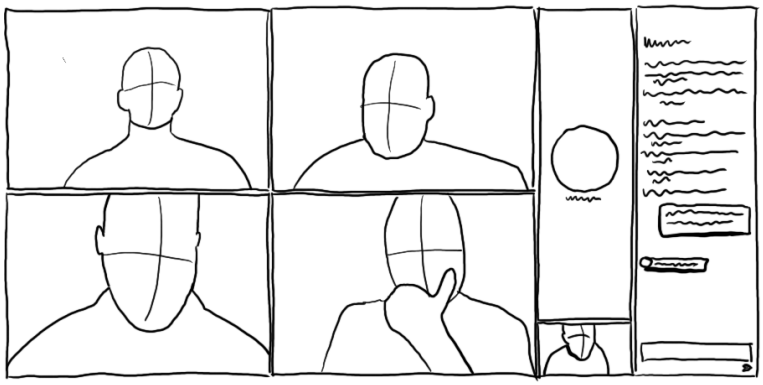


FIGURE 2.3: GSL setup for online feedback discussions

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 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. This also allowed local partners to use their language expertise to provide English summaries of discussions in local signed and spoken languages. Facilitators produced a report of the evaluation and focus group discussions following a reporting template. EUD gathered all reports and compiled them into the final deliverable, and produced an analysis based on emergent themes.

Funded by the Horizon 2020 Framework Programme of the European Union

3 APP FEEDBACK

In this evaluation, we presented participants with a prototype of the EASIER web application.[1] This version had all elements present, however, the translation function was disabled and it was not connected to the various technical components (SL recognition, translation models, avatar generation). Because the end to end translation interoperability was not stable and the maturity of translation results was volatile, we took specific measures to avoid this having a negative impact on usability testing. If participants tried to perform a translation into sign language, they were given a pre-set output sentence “Thank you for using our service” from the avatar in that sign language. The app was evaluated by deaf and hearing groups across all 5 language pairs, and was localised into all project written languages, so participants were able to select their preferred interface language. 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. (See Deliverable 8.3 **Development of client application V2** for more details on the features of the app). Appendix A contains screenshots from the app.

The app evaluation took part in 3 stages. (1) Participants were first instructed to create an account, then freely explore the app's features. (2) They were then asked to complete an online questionnaire about the app’s usability. The questionnaire was based on the traditional System Usability Scale, but presented in a bilingual format with both signed and spoken language for all 5 language pairs.[2] (3) The group then came together for a discussion which concentrated on a few major themes of the app. These themes were selected based on feedback in the interim evaluation, they were (i) Settings, (ii) Translation, (iii) Visual design, (iv) Navigation, (v) Video recording and (vi) Avatar output.

Unfortunately across several groups participants encountered difficulty with account creation and login. This was dealt with in different ways. Some facilitators prepared pre-tested user logins that participants could use if they were not able to create a login themselves. Other facilitators instructed participants to explore the app in small groups. In some cases, participants who could not log in to the app still completed the online questionnaire, but did not participate in the discussion as they had not experienced the app.

In the following chapter, we will first present results from the SUS questionnaire (section 3.1), then present the detailed feedback that arose in the focus group discussions (section 3.2), and end with some overall conclusions about the app evaluation (section 3.3).

[1] The app can be found at: https://easier-test.nuromedia.com/

[2] The questionnaire can be found at: https://questionnaire.easier-test.nuromedia.com/

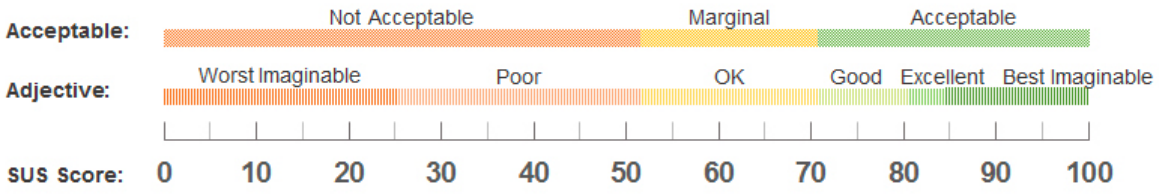
Funded by the Horizon 2020 Framework Programme of the European Union

3.1 QUESTIONNAIRE RESULTS

The questionnaire took the form of a System Usability Scale (Brooke 1996[3]), with all questions translated into the relevant signed and written languages of the groups taking part in the facilitator-led evaluation (BSL, LSF, DGS, DSGS, GSL, English, French, German, Greek). For the spoken languages, we used existing resources (e.g. original SUS, freely available translations[4]) alongside our own language expertise to produce text translations. We then used these texts as the basis to create sign language translations. When creating sign language translations, partners also consulted projects such as the ASL-SUS project (Huenerfauth, Patel and Berke, 2017[5]) to guide recording of videos in their respective language. Following recommendations from research on making questionnaires more accessible to deaf participants (Ferreiro-Lago, Pardo-Guijarro & Gutiérrez-Sigut, 2022[6]), we also offered the response scale in signed language, with a range of GIFs. In addition to the classic SUS questions and instructions, our questionnaire also included some basic demographic questions.

In total, we report on the responses of 52 participants. This number did not precisely match the number of overall participants in the evaluation, because not all participants submitted a response to the questionnaire. One major cause for this was that some evaluation sessions took place fully online. In these cases, facilitators were unable to ensure that each participant had clicked 'submit' on their questionnaire, and some participants forgot this last click, resulting in a few lost data points. In a few other cases, participants were not able to log into the app and therefore did not complete the SUS questionnaire as they felt they were unable to produce a reliable rating. Nevertheless, in other cases, despite not being able to log in, participants completed the SUS questionnaire anyway presumably rating the app based on their experiences with creating an account and logging in (this may account for some highly negative ratings).

SUS scores are typically evaluated against normative data with percentile rankings of scores (see Bangor, Kortum and Miller 2008[7]); a SUS score of 68 is considered to be average, as it stands around the 50th percentile. Several grading systems have been developed to categorise SUS scores, see Figure 3.1 below. The evaluation version of the EASIER app received a mean score of 60.1; this represents an adjective rating of 'Okay' or an acceptability rating of 'Marginally acceptable'.



[3] Brooke, J. (1996). Sus: a "quick and dirty'usability. *Usability evaluation in industry*, *189*(3), 189-194.
[4] https://github.com/ei8fdb/SUS-translations
[5] Huenerfauth, M., Patel, K., & Berke, L. (2017, October). Design and psychometric evaluation of an American Sign Language translation of the system usability scale. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 175-184).
[6] Lago, E. F., Pardo-Guijarro, M. J., & Gutiérrez-Sigut, E. (2022). Diseño de cuestionarios web en investigaciones accesibles para personas sordas mediante herramientas no estándar. *REVLES*, (4), 29-49.
[7] Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies*, *4*(3), 114-123.

Funded by the Horizon 2020 Framework Programme of the European Union

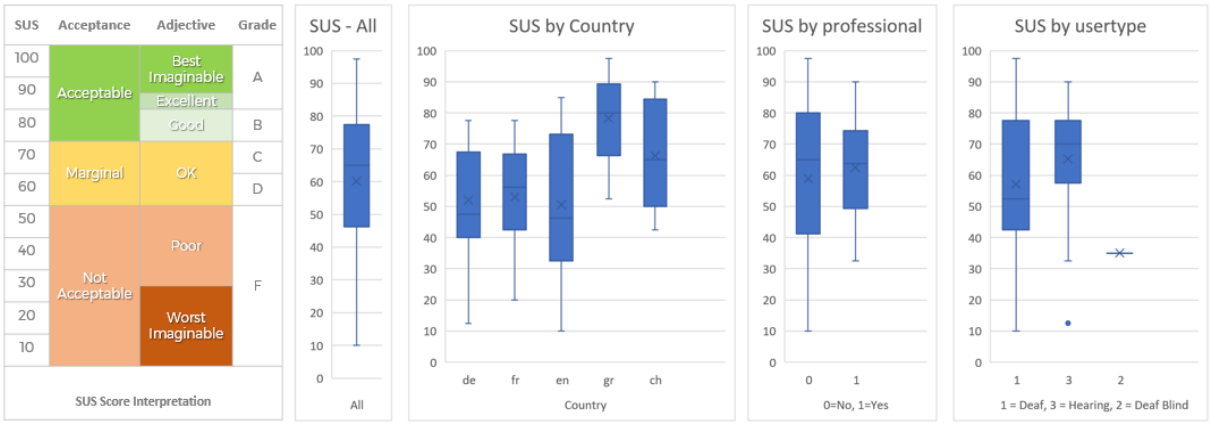
FIGURE 3.1: Scale for interpreting SUS scores (retrieved and adapted from https://measuringu.com/interpret-sus-score/)

Research has demonstrated that SUS scores allow a reliable representation of perceived insight with a relatively small sample; between 8-12 users is enough (Tullis and Stetson 2004[8]). We present the results of the SUS scale according to the different participant demographics, where all sub-groups included more than 8 participants.

Greek participants (n=12) rated SUS as most favourable with an average of 80 (range: 52.5 - 97.5), which could be considered a good app by SUS standards. The British groups (n=12) rated the lowest with a score of 46.25 however there was a large range (10-85). One reason for the very low BSL score may be that two BSL participants (one hearing and one deaf) could not log in to the app, so they just completed the questionnaire and did not join the discussion. Figure 3.2c shows the overall distribution of SUS scores per country.

Overall, scores from SL professionals (n=20) and non-professionals (n=35) were generally comparable, with a similar average score; laypeople had an average score of 61.25 while professionals rated it 58.75. Figure 3.2d shows the distribution of SUS scores by profession.

Finally, hearing participants (n=26) rated the app slightly better than deaf participants (n=28). Since there was only one deafblind participant, it is difficult to draw any conclusions from a single response, however the rating was quite low (35). Figure 3.2e shows SUS scores of deaf and hearing participants.



FIGURES 3.2a, 3.2b, 3.2c, 3.2d and 3.2e: SUS score table and the achieved scores in the final evaluations

[8] Stetson, J. N., & Tullis, T. S. (2004). A comparison of questionnaires for assessing website usability. *UPA Presentation*.

Funded by the Horizon 2020 Framework Programme of the European Union

3.2 FOCUS GROUP DISCUSSIONS

3.2.1 General feedback

3.2.1.1 Registration and login issues

During app testing, participants experienced difficulties with the registration and login process. Multiple attempts were made with various email addresses, including Gmail, and passwords, but the setup of accounts was largely unsuccessful. This resulted in negative feedback regarding the user-friendliness of the app.

Participants were unclear about the initial registration process, often entering information on the login screen instead of clicking "Registration". Additionally, there was confusion about password creation, specifically concerning password security rules and acceptable punctuation marks.

Facilitators provided tested login details to bypass registration issues. Despite this, some participants struggled with creating accounts or logging into the app after successful registration. Furthermore, the absence of a password reset option and clear confirmation of successful registration added to the participants' frustration.

The registration and login process, as the first user interaction with the app, needs substantial improvements based on the feedback received.

3.2.1.2 Browser testing vs. app testing

Participants provided mixed feedback on the app, summarising it as neither exceptionally good nor bad. They acknowledged improvements in usability and ergonomics compared to the previous version.

The fact that the evaluation was conducted via a browser rather than a downloadable app was negatively received. This aspect influenced the navigation and display, making the evaluation more challenging. Many of them had problems with the "back" function, because they were used to clicking the "back" arrow of their internet browser, which caused the application to close completely.

Participants suggested that the evaluation could potentially be better in a dedicated application version. Despite these issues, there was recognition of progress since the initial evaluation.

3.2.1.3 Status of technical development

Some participants expressed concerns about the development sequence of the app. They questioned why the app was being developed when key technical aspects such as the avatar and translation were not yet ready. The consensus was that it would be more logical to focus on these technological components first, and only evaluate the app once all elements are fully functional. The fact that they were asked to evaluate an incomplete app caused confusion among participants.

There was criticism towards projects that invest time in refining “minor details” around the translation instead of improving the core functionality. Some participants believed that the quality of the output in sign language and its legibility are more crucial for user acceptance than the app's aesthetics.

Funded by the Horizon 2020
Framework Programme of the European Union

3.2.2 Settings

3.2.2.1 Preferred settings

Participants had mixed feelings about the option to choose whether or not to use preferred settings. Some disliked the requirement to make this choice and suggested that the app should simply remember the last used settings. Others, however, favoured the ability to set preferred input and output settings.

One participant proposed that upon first opening the app, users should be asked about their preferred settings, rather than having to navigate through all settings. The app should also ask each time it's reopened whether the preferred settings should be used.

Deaf users reported confusion about what "Use favourite settings" referred to on the app's home page. It was unclear whether this applied to communication mode and language choices, avatar settings, contrast settings, or all of these. This led to some difficulty in reusing language choices set in the settings menu on the translation page, as they had not ticked the "use my favourite settings" box.

FIGURE 3.3: Use favourite settings in the French version of the app

Participants reiterated a suggestion from the midterm evaluation for the ability to save multiple profiles to quickly adapt to different situations. They found the current settings process time-consuming and complicated.

3.2.2.2 Difference between settings and profile unclear

Some users expressed confusion about the distinction between the "settings" and "profile" functionalities. They did not immediately understand the difference between the two. For them, "settings" should be within "profile," as the term "settings" implies app settings e.g. the dark mode, not language choice as they were of the opinion that such settings would be under “profile”.

This discussion echoes a previous conversation during the mid-term focus group, where users compared this application to others they regularly use. In those applications, communication preferences are typically linked to "profile" settings.

3.2.2.3 Location of settings

Participants were unsure about the location of certain settings and suggested adding more information to the main settings page. They found it cumbersome to open each setting option individually to find what they were looking for. They also proposed adding information to the app's landing page about the preferred settings, such as a memory aid for the current input and output language. Some also confused the language settings for the app and those for the translation.

Funded by the Horizon 2020
Framework Programme of the European Union

While participants appreciated the range of customizable settings (like the background colour and avatar clothing, script size, and dark mode), not all of them found these settings without assistance. They expressed dissatisfaction with the arrangement of the settings and their locations. They suggested that the avatar settings, font size, and dark mode settings should be on the same page, as that is where they would expect them to be.

This feedback indicates a need for more intuitive organisation and clearer guidance within the settings options to enhance user accessibility and ease of use.

3.2.2.4 Avatar settings

Participants appreciated the ability to change the avatar's settings but had varying opinions on the selection choices. While one participant desired a full palette of colours, another preferred the simplicity of only three options, suggesting that more choices could be overwhelming and time-consuming.

The option to individually adjust the contrast of the background and clothes was well-received. However, participants criticised the lack of light colour options for clothing and the absence of skin colour customization for the avatar.

Despite an icon indicating gender options, participants noted it was not possible to change the avatar's gender. Besides these points, the settings were generally deemed straightforward by participants.

3.2.2.5 Suggested additional settings

Participants suggested additional settings for improving the app's output. Specifically, a volume adjustment feature was proposed for spoken output, which would be helpful depending on the surrounding noise levels where the app is being used.

Furthermore, participants recommended features for adjusting the output display of the avatar, including changing the speed and zooming in or out. This would allow users to tailor the signing to their comfort level. These suggestions indicate that users desire more control over the app's output settings for a more personalised experience.

3.2.3 Translation flow

3.2.3.1 Interaction flow

Participants acknowledged improvements in the linear progression of the application, but felt constrained by its "one-way" interaction. The current structure does not allow for a smooth back-and-forth dialogue, as switching language combinations requires returning to the settings. This constrains the shape of the interactions between the interlocutors and does not allow a smooth exchange.

The absence of a reverse icon for quickly swapping languages was noted, with participants comparing the app's interaction to consecutive translation rather than the simultaneous translation provided by real interpreters. One suggestion was to have input and output language choices on the same page to simplify the translation process and provide a clearer overview. Once choices are saved, a page displaying the final language combinations would be shown.

However, one participant expressed dissatisfaction with the time required to explore features, use the video application, and obtain desired translations. This individual found the visual

Funded by the Horizon 2020 Framework Programme of the European Union

design plain and less appealing than other services, and suggested the app may not be user-friendly for those less confident with technology, such as older individuals or those with complex needs. The app's usability was thus identified as a potential barrier to its adoption.

3.2.3.2 Input and output options

Participants appreciated the variety of input and output options for the translation feature, which included speech, text, and signing. The process of translation was mostly considered straightforward, but some suggested simplifying the steps to change the output language, which currently requires more steps than changing the input language. They expressed dissatisfaction with the dissociation of input and output default settings. They preferred having input and output choices presented on a single page. This format would reinforce the application's bidirectional usability.

Participants were frustrated by having to restart the entire translation process when they only wanted to change a language or the direction of the translation, as there is currently no simple way to change the direction of translation. For use cases where they are in a dialogue, they suggested implementing an option to change the direction of the input and output languages with one click, and adding options to change the language within the same screen. These suggestions reflect a desire for a more streamlined and user-friendly approach to changing languages and translation directions within the app.

However, some participants were not sure what they should evaluate as the translation is only a mock-up. This limited their ability to evaluate the feature effectively. Although the process of translation was clear, the participants found it neither satisfactory nor successful due to these technical limitations.

3.2.3.3 Input deleted upon closing the app

Another participant did not like that the input text for translation is deleted when closing the app. They suggested that the text entered is saved in the input window and is still there when reopening the app, or that one can choose if the input should be saved or not.

3.2.3.4 Doubts during the flow

Some participants raised concerns about the clarity of the translation process within the app.

Firstly, it was not evident when the translation had been completed. Some participants liked the "translation in progress" display during input processing. This confusion was partly due to the placeholder text "text output" used in the absence of integrated translation functionality. Participants suggested a clear message such as "translation finished" to indicate the completion of the translation.

Funded by the Horizon 2020
Framework Programme of the European Union

FIGURE 3.4: "Translation in progress" screen in German

Secondly, the button to initiate translation was not clearly labelled. One participant expected a button labelled "translation", but found that the actual command was "start". It was suggested that the button be labelled "translate" or "start translation" for greater clarity. One participant stated that she did not find where to do a translation and was thus unable to test this part of the app. These issues highlight the need for clearer instructions and labelling within the translation process.

3.2.4 Visual design

3.2.4.1 Background settings

Some of them said that it was easy and straightforward to find, adjust and save the settings. However, they said it would be good to see additional options like changing the colours of the background to suit people with additional needs e.g. visual sight, Autism Spectrum Disorder (ASD) and anxiety.

3.2.4.2 Icon choice

While most icons were considered self-explanatory, participants raised a few concerns. They criticised the use of the same flag for different languages and suggested the addition of small symbols to differentiate signed languages from spoken ones, e.g. hands for signed languages and lips for spoken ones.

The icon for the avatar was seen as misleading, as it is commonly recognized as a standard icon for user profiles or accounts. Participants suggested using the face or outline of the avatar instead. They compared this icon unfavourably to the one used in iLex[9]. These comments highlight the importance of intuitive and universally recognized iconography in the app's design.

[9] https://www.sign-lang.uni-hamburg.de/ilex/

Funded by the Horizon 2020
Framework Programme of the European Union

FIGURE 3.5: iLex icon

3.2.4.3 Dark mode

Participants appreciated the inclusion of a dark mode in the app, noting that black and yellow were expected colours for this feature. However, they expressed interest in a second type of dark mode, not specifically designed for high contrast for deaf-blind users, but simply for a darker layout. They mentioned that many of them use darker layouts on their personal devices and enjoy apps that automatically switch to dark mode when the device setting is adjusted. For this type of dark mode, they preferred colours like grey and white instead of yellow.

FIGURE 3.6: Dark mode as it looks in the EASIER app

3.2.4.4 Layout of output screen

Participants expressed dissatisfaction with the layout of the output screen. They suggested displaying the input and output on the same screen. While they acknowledged that placing them side by side (as done in e.g. Google Translate) may not be feasible on a phone, they proposed arranging them vertically instead.

They also recommended removing the 'ears' above the input/output window displaying the input and output language, suggesting smaller script labels as a replacement.

Funded by the Horizon 2020
Framework Programme of the European Union

FIGURE 3.7: The 'ears' with input and output language

Participants were confused by the inconsistent placement of the return button, which appears at the bottom of the page for the output screen but not on other screens. They advocated for more coherence in the layout.

One participant proposed the ability to scroll up and down within text display windows. These suggestions reflect a desire for a more intuitive and consistent layout design on the output screen.

3.2.4.5 Design choices

Some evaluators noted an improvement in the aesthetics of the application, finding it attractive, simple and clear compared to the first evaluation. They appreciated that the interface is now more responsive.

However, others found the initial interface plain and uninviting, suggesting there should be e.g. a visual display of someone signing "welcome" in different languages. The small font size also contributed to the difficulty in understanding what they were meant to do. One participant suggested making the colour scheme 'more friendly' as they found the combination of red and grey 'a little bit old fashioned'.

The world map on the screen for general settings was confusing to some participants, as it seemed to serve no purpose and the colour coding of the countries lacked assigned meaning.

FIGURE 3.8: The world map image which was named "confusing"

These feedback points indicate that there is a wide spectrum of opinions on the design choices for the EASIER app, ranging from positive to negative.

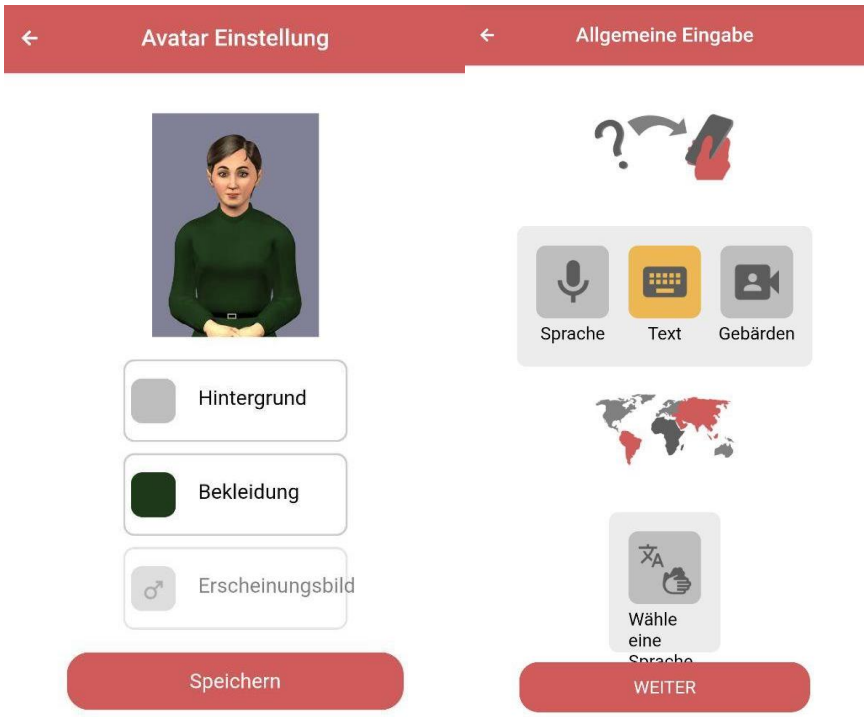
3.2.4.6 App starting up in English

The default language of the app upon the first time opening was English. Some pointed out that this didn't bother them but that some people who didn't speak English might find themselves stuck.

Funded by the Horizon 2020
Framework Programme of the European Union

3.2.4.7 Visible localization issues

In some browsers there was a problem with the display of the text, going outside of visual elements.



FIGURES 3.9a and 3.9b: Issues with text display in German

3.2.5 Navigation

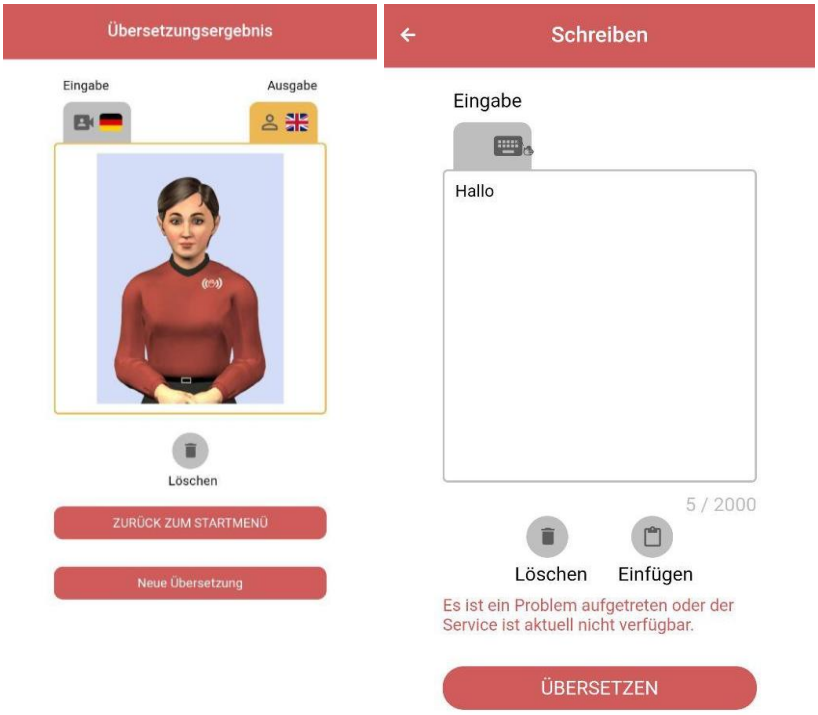
3.2.5.1 Complex navigation

Participants found navigating the app too complicated. They weren't sure what the 'Start' button would actually start and didn't understand that they needed to choose all settings before starting the translation. They described the journey from start to translation as 'too long' and suggested having a drop-down menu for selecting input and output languages on the output screen, as the current solution was confusing.

The app was felt to have too many buttons and lacked clear signposting for different options. Some participants had issues with 'going back,' which was not intuitive. The navigation was described as 'long distance,' indicating too many steps or surfaces that could be combined into one.

The placement of the return button was inconsistent, sometimes being at the top and other times at the bottom of the screen. This inconsistency, coupled with some technical issues like scrolling not working properly, hindered the user experience.

Funded by the Horizon 2020
Framework Programme of the European Union



FIGURES 3.10a and 3.10b: Differences between return button placements

There was also confusion about the language selection. Participants were unclear about the need to select a language combination by clicking on the grey tab. This was further complicated by the colour coding convention where grey often indicates a 'deactivated' option.

The French terminology used in the app did not reflect the logic of their language (LSF), leading to further confusion. These comments highlight a need for clearer, more intuitive navigation and language selection processes within the app.

3.2.5.2 Confusing terminology

Participants in the French evaluation groups appreciated that the application was available in French upon logging in. However, the French terminology for various app functions led to confusion.

The terms for input and output, "entrée générale" and "sortie générale," were particularly problematic. Only one participant, who works in audiovisual editing, understood these terms immediately. For others not familiar with this terminology, it was not evident, especially when associated with certain images. The terms did not suggest the notion of translation and were commonly associated with general entrance or exit of a building or room.

FIGURE 3.11: Examples of confusing terminology in French

The terms "mode de saisie" and "mode de sortie" also caused confusion as they were unrelated to the concept of translation. Given the complex approach French deaf people often

Funded by the Horizon 2020 Framework Programme of the European Union

have towards reading in French, the polysemous word "mode" was associated by some participants with fashion, rather than its intended meaning.

There was another suggestion to better label the “start” button as “start translation” or “translate,” and to include a “translation finished” message for clarity.

One participant expressed dissatisfaction with the German wording “Allgemeine Ausgabe/Eingabe” and suggested using “Eingabesprache/Ausgabesprache” instead. These responses indicate a need for clearer, better localised terminology within the app.

3.2.5.3 Drop-down menu

Some were confused by the order of the drop-down menu. For them, "disconnect" should be at the very bottom of the menu. They pointed out that in all the applications and sites that used a connection, this function was always at the very bottom.

FIGURE 3.12: Drop-down menu

3.2.5.4 No clear instructions

One participant found the settings particularly confusing to navigate, noting a lack of clear instructions on how to adjust them to their preference. They suggested that this aspect needs further refinement.

Another participant expressed initial uncertainty about what to do when asked to amend the settings. They first saw the camera icon and thought they needed to click on it and sign to start the app, but they quickly realised the actual requirements.

These comments highlight a need for clearer guidance in the application's navigation.

3.2.5.5 Menu in home page after login

Participants raised concerns about the initial display of the app after login, specifically regarding the drop-down menu. They reported that upon reaching the home page, the menu at the top left was already open and some found it difficult to close. The open drop-down menu blocking parts of the screen upon app startup was identified as a flaw by the participants.

3.2.6 Video recording

3.2.6.1 Video input layout

Funded by the Horizon 2020 Framework Programme of the European Union

Participants expressed dissatisfaction with the layout of the video input window. They felt that too much space was left unused and suggested reducing or removing the frame around the recording window. A full-screen recording window was also proposed, with participants noting that the small current window was straining to their eyes, or that the video was not mirrored, which is confusing when signing.

FIGURE 3.13: Video input screen

Some commented that they did not find the video/audio recording controls to be intuitive. For example, one said: “So it was like: click on the record symbol which was in the middle of the screen. And I thought that was recording. But actually, what I needed to do was then also press the white dot down in the left corner. This wasn't obvious. I’m used to seeing the record symbol in red as well.”

Suggestions also included enlarging the recording button, placing the recording button on a location that is easily reachable with one’s thumb and adding a countdown before the recording starts for a better user experience. One participant expressed a desire for "a better haptic experience" when recording a video.

This highlights a need for a more user-friendly design and functionality in the video input layout. Participants also recommended adopting common designs known from other apps.

Funded by the Horizon 2020 Framework Programme of the European Union

FIGURE 3.14: Instagram video/picture input screen which was named as an example to follow

One participant reported that the interface for the audio recording was mirror-inverted.

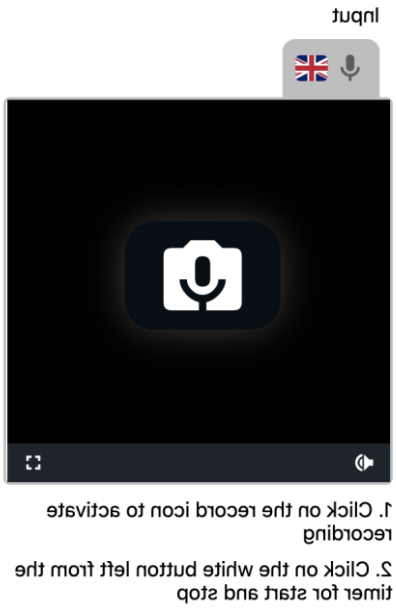


FIGURE 3.15: Mirror-inverted recording screen

3.2.6.2 Input variation

Participants raised questions about the flexibility of the app's input methods.

One query was about the recording background. It was unclear whether a uniform backdrop was necessary for the app to function correctly, or if recording in front of a non-uniform background like a bookshelf would still yield accurate results. What are the minimal requirements for colour of clothing, background and lighting?

Funded by the Horizon 2020 Framework Programme of the European Union

A concern about one-handed signing was raised by a user familiar with using video relay services on their mobile phone. They pointed out that there are situations where it's not possible to put the phone down, necessitating signing with one hand. The question was whether the app could recognize sign language when only one hand is used. This highlights a potential need for the app to accommodate various signing conditions and environments and to inform users which those are.

3.2.6.3 Voice input requires video recording

Participants were confused that voice input required a video recording and not only a voice recording. They would prefer voice recording only.

3.2.6.4 Video input duration

Questions were raised about the limitation of video duration to 1 minute. Participants were unsure whether this restriction was due to storage concerns or the length of translation the app could handle.

3.2.7 Avatar output

3.2.7.1 Diversity of avatars

Participants expressed a desire for more extensive avatar customization options. They wished for a broader selection of colours for the avatar's clothing and background. They also expressed interest in non-white avatars and suggested that in such cases, clothing should be lighter to improve contrast. One participant reported on a presentation (made by EASIER project members[10]) at the SLTAT workshop in Rhodes underscoring the need for diversity in avatar representation.

The participants were interested in altering Paula's appearance. Some wished for complete customization akin to The Sims game, while others preferred limited choices to avoid potential annoyance. Another participant expressed a preference for non-human avatars, citing concerns about the future over-humanization of avatars. These responses indicate a desire for diverse and extensive avatar customization options.

3.2.7.2 Video output

Feedback on the avatar output was limited due to the availability of only one video. However, participants deemed the display acceptable. They suggested that the screen space could be better utilised by enlarging the avatar window at the sides, but noted a preference against a full-screen avatar.

They also suggested bringing buttons closer together to prevent the need for scrolling to access buttons located below the avatar window. It's worth noting that the visibility of these buttons may vary depending on the phone model used. These suggestions indicate a desire for a more efficient and user-friendly layout in the avatar output display.

3.2.7.3 Original colour settings unavailable

[10] Kopf, M., Omardeen, R., & Van Landuyt, D. (2023, June). Representation matters: The case for diversifying sign language avatars. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)* (pp. 1-5). IEEE.

Funded by the Horizon 2020 Framework Programme of the European Union

Participants noticed that when the settings for the avatar are changed the original colour settings disappear. When starting the app, the avatar is wearing a red sweater, this setting is not available and once changed cannot be chosen again.

3.2.8 Archive

3.2.8.1 Sort and search

While participants generally appreciated the archive and its date-based sorting, they expressed a desire for more diverse sorting options. They disliked the naming convention of "archive..." for translations in the archive, suggesting it was redundant and unhelpful. They proposed being able to name files themselves or display the beginning of the translated text.

This led to a discussion about what would be shown for signed input, with a small video preview being suggested. One participant even suggested showing both input and output, but this idea was not explored further. Questions were also raised about archiving dialogues and whether it would be possible to save them as a bundle or if each input would need to be saved separately.

Participants liked the concept of the archive but wished for the ability to mark favourites or add tags to make translations easier to find later. They suggested use cases such as tagging texts with different topics like 'university texts' or 'every-day texts'.

Overall, participants expressed individual needs that varied widely but all agreed on the necessity for sensible sorting or a search function, preferably with integrated sign recognition. This feedback reveals a desire for enhanced navigational and organisational features within the archive.

3.2.8.2 Usefulness

Participants had differing opinions on the usefulness of the archive feature. One participant saw no use for it, while others proposed various use cases to explain its utility. These included saving frequently used sentences for asking for regular facilities (e.g., WC), travel directions, or information that one regularly presents and can prepare beforehand.

There was some confusion about the purpose of the archives. Users did not immediately understand its function and required explanation that it was meant for archiving frequently used phrases. One participant mentioned initially visiting the Archive expecting to find content there and felt disoriented when it was empty.

The "delete" option in the Archive also caused some confusion. Two participants wondered what they needed to delete and subsequently skipped the option without trying it.

One participant expressed satisfaction with the overall navigation but found accessing the archive unintuitive, as it required returning to the home screen first. This feedback suggests a need for clearer instructions and more intuitive design around the Archive feature.

3.2.9 Other

3.2.9.1 After the project

Some participants expressed concerns about the app's sustainability post-project, specifically in regards to technical support and regular updates. They questioned the long-term durability of such apps once the project period concludes.

Funded by the Horizon 2020 Framework Programme of the European Union

There was a broader discussion on the nature of short-term research projects, with critique towards topics being dropped after the end of these projects. Participants emphasised the importance of collaboration between the IT sector and sign linguistics sector in these projects.

The role of both hearing and deaf researchers was also discussed, highlighting the need for diverse perspectives in the development and continuation of such initiatives. One commented that the app was designed by hearing people, that “it was obvious that it’s not intuitive or deaf friendly”.

3.2.9.2 Privacy and GDPR

Participants raised concerns about data privacy and storage in relation to the app's translation feature. It was unclear whether the app retained video recordings of users' signing for translation purposes, leading to questions about the app's privacy policy and GDPR compliance.

Some participants saw an advantage in recording the signing as it allows verification of the intended sign. However, they expressed discomfort with the idea of automatic saving of every translation. Suggestions included a switch to enable/disable automatic saving of translations, considering cases where users might want to remain anonymous.

Questions were also raised about the location of stored material. Participants expressed concern about local storage potentially filling up disk space and cloud storage potentially compromising data privacy.

Notably, some participants reported their phones indicated that the camera continued to be used even after closing the web browser. This raises further privacy and data usage concerns that need to be addressed.

3.2.9.3 User manual

There was a consensus on the need for a user guide or tutorial to assist new users. Upon initial login, some evaluators were unsure of the next steps and went directly to the translation page without accessing settings. A popular suggestion was to include a tutorial during the first login to familiarise users with the application and its various functions.

Participants also proposed the inclusion of use cases when a user first accesses the app, explaining its potential uses, its capabilities and how to utilise them. Additionally, step-by-step instructions for initial setup, such as visiting the settings first, were recommended.

These suggestions indicate a desire for more explicit guidance within the app, particularly for first-time users. This could be in the form of an optional user guide or an introductory tutorial. It was also mentioned that FAQ or contact information for technical support is missing, in case a technical problem occurs.

3.2.9.4 Use cases

Use cases that came up during a discussion were: a dialogue between a non-signer and signer, texts that are used on a regular basis (e. g. a handbook for a technical widget, questions for directions). One person thought that this app is not designed to be used in dialogue situations.

Some participants asked if it could be possible to choose any combination of languages, including from one sign language to another which was regarded as useful.

Funded by the Horizon 2020
Framework Programme of the European Union

3.2.10 Out of scope suggestions or questions

3.2.10.1 Interconnecting phones

Participants suggested the addition of an interconnectivity feature for phones, to improve the usability of the app during dialogue. The conversation on this topic was initiated by a participant who expressed discomfort with the idea of passing their phone back and forth with a conversation partner.

The consensus in this specific conversation was that the app should include a QR code for easy recommendation and download. Following this, the two phones should be able to connect in such a way that both users can see each other's translations. This would facilitate a smoother dialogue without the need to share one device.

As a reference, participants pointed out that the app Ava has a similar feature.

3.2.10.2 Connect app to screens

A further suggestion for an additional function was the ability to connect the app to screens on public places. For example, to be able to show the information on the screens on a train station’s track live on one’s phone. This way one could see the information without having to walk all the way up to the next screen.

3.2.10.3 3D avatar projection

Another recommendation for the future was the display of the translation output in the form of a 3D avatar projection into the room. This would have the comfort that the user can look at it from different perspectives for better understanding.

3.2.10.4 Offline availability

One question that arose was if the app will be available in an online and offline version. If there is an offline version participants wondered how much disk space this would use on their phones.

3.2.10.5 Scanning text

An additional requested feature that was named, is the option to scan written text as an input for translations. Named use cases for scanning written text were menus in restaurants and the running texts in public transport.

Funded by the Horizon 2020
Framework Programme of the European Union

3.3 CONCLUSION

Much like the interim evaluation study, the app evaluation generated a lot of engaged feedback from end users. On average, participants rated the prototype version of the web app as “okay” on system usability; nevertheless, there was a range of ratings with certain countries rating the app higher than others and hearing users rating the app slightly better than deaf users. While these results provide a useful benchmark for future work, it should be noted that participants evaluated a prototype lacking some key functionality, namely translation. With an integrated translation function, we anticipate users will produce a higher SUS rating for the app.

The qualitative round of feedback provided useful information on the strengths and weaknesses of the app, providing a roadmap for the fine tuning to be done during the remainder of the project. In the last phase of the project the outcome of the evaluation will be prioritised and in the app UI will be refined. Focus will be on the simplification of click flows and enhancement of the browser integration. While all feedback points will likely not be able to be incorporated, they also provide useful and thoughtful starting points for future work on designing translation apps for deaf users.

Funded by the Horizon 2020
Framework Programme of the European Union

4 AVATAR FEEDBACK

The avatar evaluation took the form of a questionnaire, followed by a group discussion. This questionnaire was developed by ATHENA to collect feedback on the avatar from deaf community members. The questionnaire was prepared for the four sign languages for which the avatar is currently available: GSL, DGS, DSGS and LSF. Thus, eight groups (deaf and hearing from each of the four languages) completed this part of the evaluation. Questionnaires for each language pair were bilingual with both text and sign language and contained signed instructions for navigating each page (see Deliverable 2.2 **Final sign language avatar** for more information about the development of the questionnaire). Appendix B contains screenshots from the questionnaire.

The app evaluation took part in two stages. (1) First, participants were given the link and directed to complete the questionnaire.[11] 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. Here there was a strong focus on presenting old and new versions of the avatar to better understand how end-users viewed the changes (for details on the differences between avatar versions, please refer to Deliverable 2.2). At the end of the questionnaire, participants were invited to record a video providing their feedback. (2) After completing the questionnaire, participants returned to the group setting 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.

[11] The questionnaire can be found at: https://sign.ilsp.gr/slt-eval-2

Funded by the Horizon 2020
Framework Programme of the European Union

4.1 QUESTIONNAIRE RESULTS

The questionnaire first asked for background demographic information about participants, with respect to their age, gender, age and context of sign language acquisition and sign language proficiency.

Participants saw a series of screens for each animation. On the first screen (Figure 4.1), they viewed a video of a human signer, which was used to create the avatar animations. Then on the next screen (Figure 4.2) they viewed two avatar animations side by side and were asked a series of questions. First, they had to identify which of the two avatar animations was better. Then they were asked to rate both videos; first on a five-point scale ranging from “Very good” to “Bad”. On the third screen (Figure 4.3) they viewed the two animations side by side again and were asked to rate them each on (1) facial expressions and head movements, (2) mouth movements, (3) hands and body, and (4) the intelligibility of the signing. All were rated on a five-point scale ranging from “Very good” to “Bad”.

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 app.

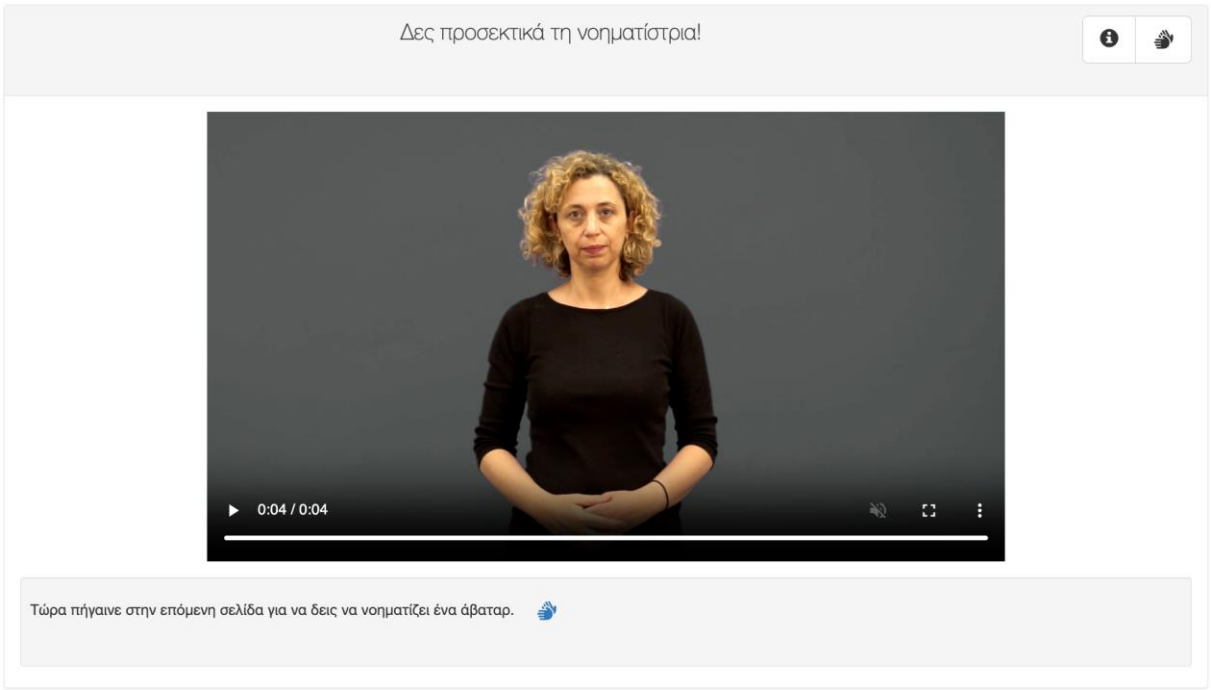


FIGURE 4.1: First step: Performance of signed evaluation material by a human signer (in the GSL questionnaire in the figure)

Funded by the Horizon 2020
Framework Programme of the European Union

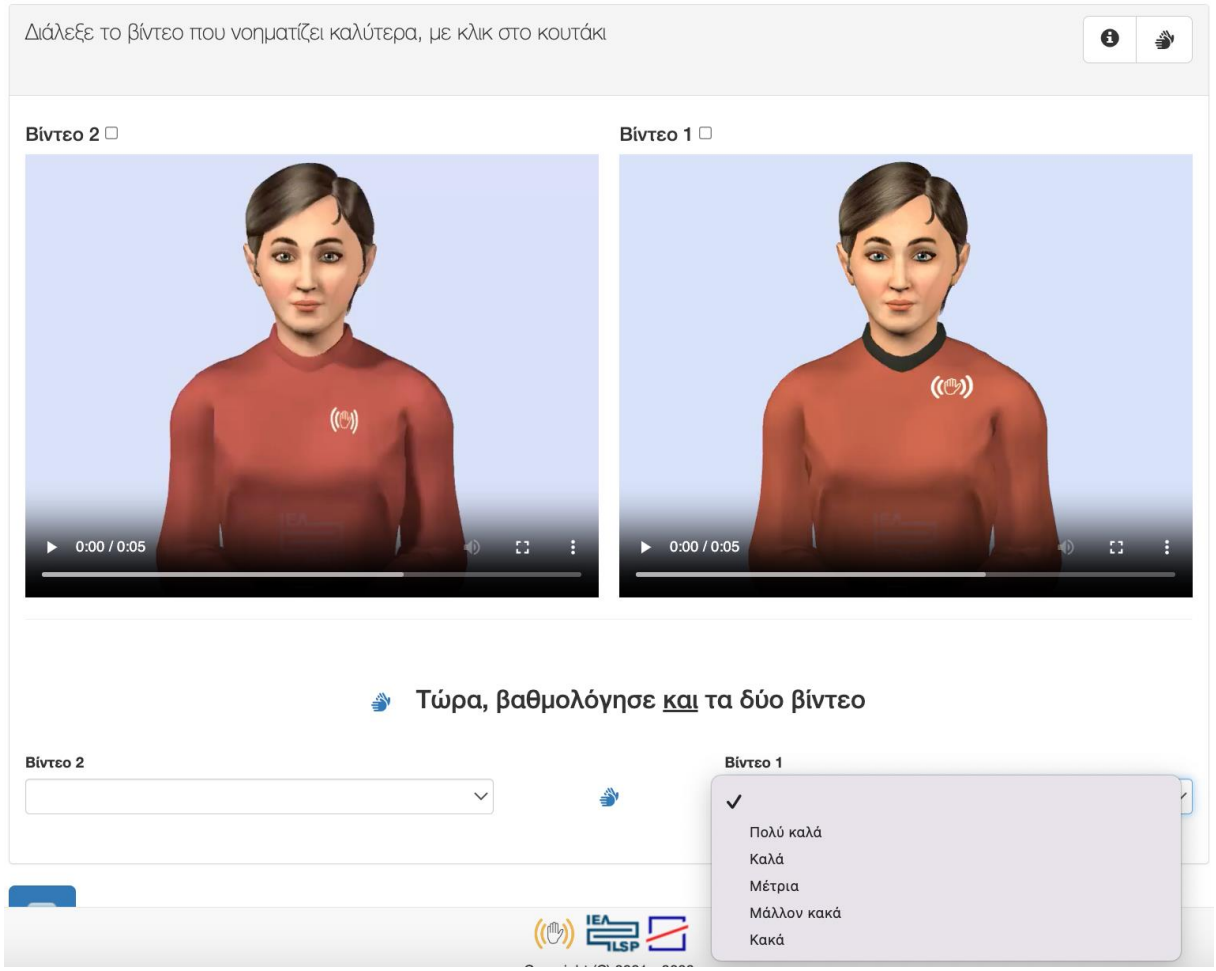


FIGURE 4.2: Second step: Choice of preferred avatar performance and rating of both avatars

Funded by the Horizon 2020
Framework Programme of the European Union

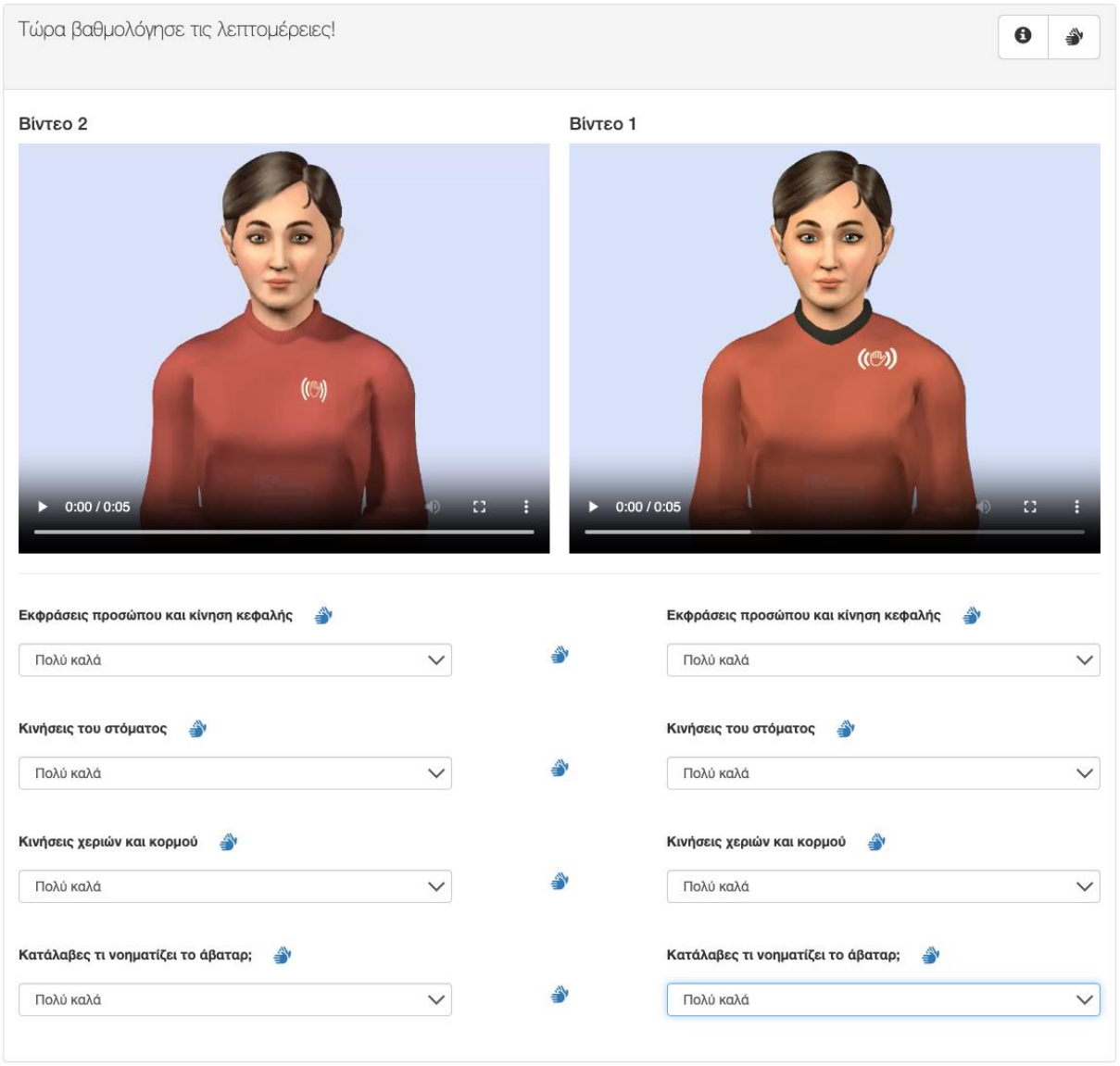


FIGURE 4.3: Third step: Rating of signing performance details

For a clearer view of end user groups' ratings of the avatar performance, we present next the overall ratings for avatar readability and acceptance per language and overall, for the current and previous version of the signing avatar. All results are presented via the same colour code for the groups of deaf (green), hearing (orange) and overall figures (brown). DG refers to the Deaf Group and HG refers to the Hearing Group. On the rating scale 1=Very well, 2=Well, 3=So-so, 4=Rather bad, 5=Bad; therefore, lower scores are considered better while higher scores are considered worse.

Funded by the Horizon 2020
Framework Programme of the European Union

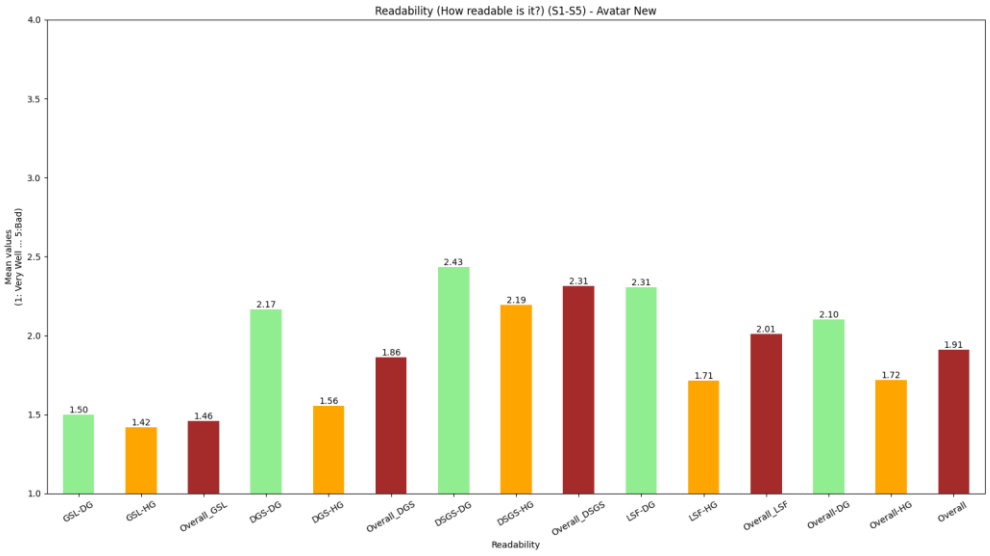


FIGURE 4.4: Readability scores for the current avatar (Avatar New)

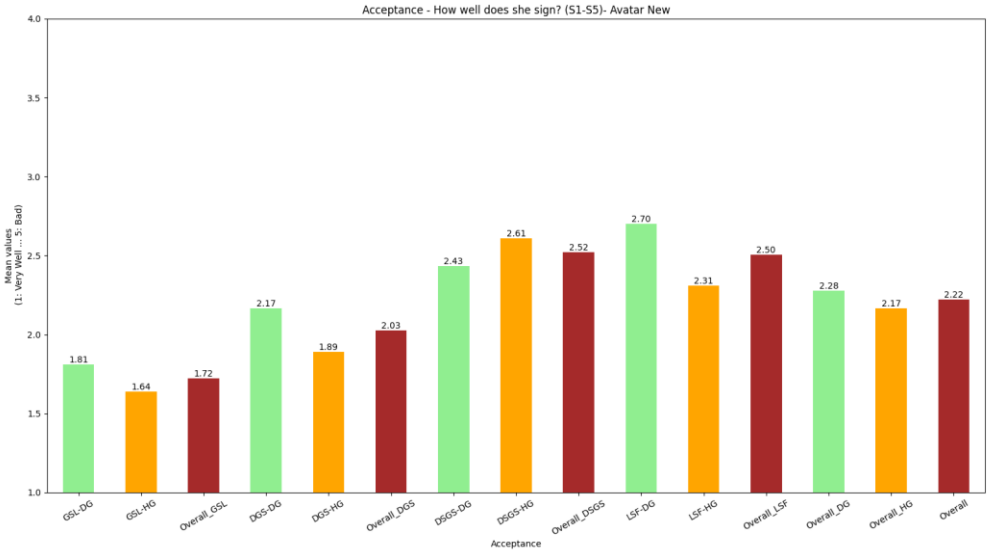


FIGURE 4.5: Acceptance scores for the current avatar (Avatar New)

As a general comment, GSL users provided the best scores regarding both acceptance (an average 1.72) and readability (average 1.46). The worst scores came from the DSGS and LSF groups with average acceptance 2.52 and 2.50 and average readability 2.31 and 2.01 respectively. Similarly, the hearing group systematically provided better scores than the deaf group in almost all cases. A most interesting finding is the difference between readability and acceptability scores, which also complies with the variety of the comments received during the discussions on user evaluation experience.

Funded by the Horizon 2020
Framework Programme of the European Union

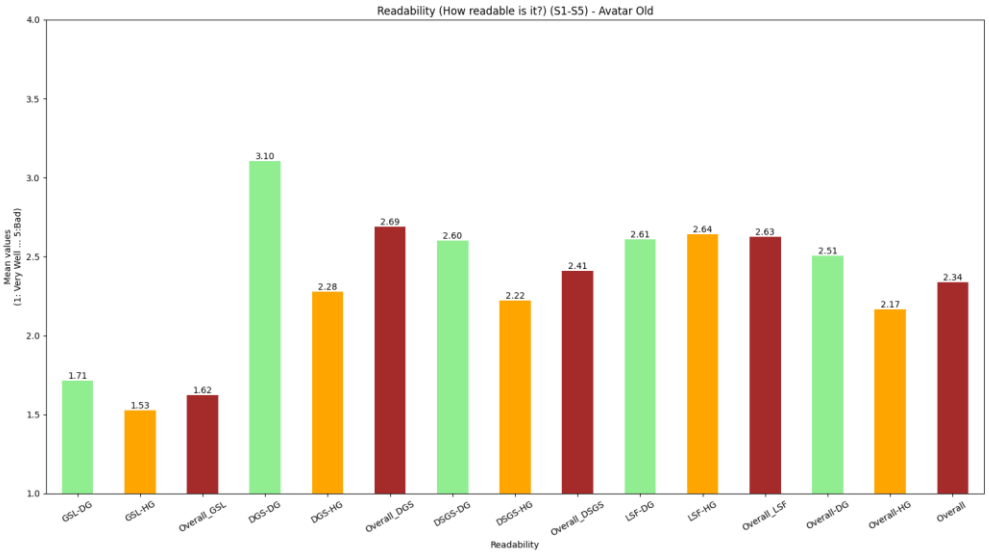


FIGURE 4.6: Readability scores for the previous avatar version (Avatar Old)

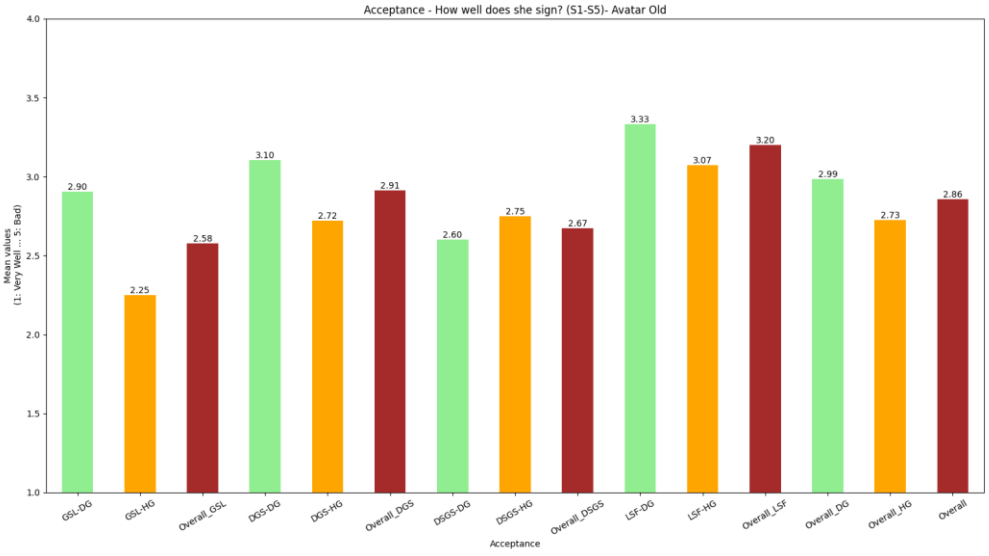


FIGURE 4.7: Acceptance scores for the previous avatar version (Avatar Old)

Overall improvement in avatar acceptance and readability is also noticed from the comparison of evaluation results (in Table 2) between the previous and current versions of the EASIER avatar as depicted in the end user scores for the previous version of the avatar in figures 4.6 and 4.7.

TABLE 2: Comparison between old and new avatars (the lower, the better)

Funded by the Horizon 2020 Framework Programme of the European Union

	Readability: How readable is it?														
Old	1.71	1.53	1.62	3.10	2.28	2.69	2.60	2.22	2.41	2.61	2.64	2.63	2.51	2.17	2.34
New	1.50	1.42	1.46	2.17	1.56	1.86	2.43	2.19	2.31	2.31	1.71	2.01	2.10	1.72	1.91
	Acceptance: How well does she sign?														
Old	2.90	2.25	2.58	3.10	2.72	2.91	2.60	2.75	2.67	3.33	3.07	3.20	2.99	2.73	2.86
New	1.81	1.64	1.72	2.17	1.89	2.03	2.43	2.61	2.52	2.70	2.31	2.50	2.28	2.17	2.22
	GSL DG	GSL HG	GSL All	DGS DG	DGS HG	DGS All	DSGS DG	DSGS HG	DSGS All	LSF DG	LSF HG	LSF All	All DG	All HG	Overall

Funded by the Horizon 2020
Framework Programme of the European Union

4.2 FOCUS GROUP DISCUSSIONS

This section presents qualitative data gathered during the focus group discussion that followed the completion of the questionnaire.

4.2.1 Overall appearance and animation

Several groups noted global improvements to the avatar; for example, DGS hearing participants noted that the new version of Paula is better animated than the old version, deaf GSL participants appreciated the significant improvements. LSF participants were pleased with overall improvements in legibility and pointed out that when compared to other avatars they are familiar with in public spaces such as an avatar designed for train stations in France, [12] Paula is significantly better. DGS participants were also pleased that there were plans for a male and non-binary avatar, and suggested Paul and Pauli (respectively) as fitting names for these avatars.

DGS participants appreciated the good contrast between eyebrows and face colour as well as the improvement of less intense but still visible shadows on Paula's body. They also suggested that Paula's sleeves were a little too baggy to properly see the shape of her arms, creating a strange boxy shape of the elbow; they would prefer tighter sleeves and a more natural looking elbow. LSF participants suggested that better animation cues on the hands could support users more quickly identify hand configuration, for example more colour contrast between nails and skin.

While the French still found Paula to be overall robotic, not all participants viewed his negatively; some thought it was reassuring as there was no doubt of interacting with an avatar, while others found this made Paula appear cold and difficult to bond with. Nevertheless, deaf DGS participants pointed out that despite Paula's shortcomings (intonation, fluency, non-manuals) they would easier be able to overlook these issues when using Paula for short utterances as opposed to longer periods of watching (e.g. 2-3 minutes).

The DGS hearing group had several comments concerning Paula's presentation in the app. They noted that while Paula's eye gaze suggests looking down at the user, the camera perspective suggested the user looking down on Paula. They suggested a change in perspective to address the perceived mismatch between gaze, head position and camera angle. They also suggested an option to zoom out and spin Paula around to view signing from different angles (this could present the opportunity for fun 'Easter eggs' such as Paula wearing fuzzy slippers on her feet).

4.2.2 Prosody

Both LSF deaf and GSL hearing groups noted an improvement in Paula's fluency when compared to the previous evaluation. They commented that better rhythm, more fluid transitional movement between signs and improved sentence level prosody made Paula more comfortable to watch.

4.2.3 Manual signing

[12] https://www.sourds.net/2010/11/26/jade-arrive-dans-les-gares/

Funded by the Horizon 2020 Framework Programme of the European Union

Several positives were noted. The French were unanimously pleased with the improved precision of hand configurations. Greek hearing participants also praised the improvement in Paula's hand motion, and appreciated the slight rotation of specific signs to improve occlusion.

However, other aspects of manual signing were identified as places for improvement. Greek deaf participants pointed out that Paula's movements could be made to look more natural by adding relaxation of hands between movements. French participants noted that Paula's signing space is too small, making it more difficult to understand her. As one French deaf participant described it "*It's as if you are too close to her, which makes it uncomfortable to listen*". To improve this, they suggest a 3/4 shot of Paula's body, and a larger signing space.

4.2.4 Non-manuals

Several groups noted overall improvements in non-manuals; both Greek groups, as well as DSGS hearing, and DGS deaf groups all remarked positively on Paula's non-manuals. They particularly noticed improvements in the upper body, head and face, including eyebrows and forehead.

The DGS hearing group appreciated eyebrow movement specifically, and some also commented positively on the wrinkles around Paula's mouth and cheeks as helpful. On the other hand, some participants found these wrinkles were over exaggerated to the point of distraction, and 'unnaturally visually salient', not matching the rest of the animation, for example, with the sign VERSTEHEN (Figure 4.8). One DGS hearing participant mentioned that while she asked for more non-manual features in the interim evaluation, now that they are more emphasised, she thinks it is too much. One LSF participant felt that the wrinkles made Paula feel 'colder' while another commented that they "*made the avatar look 20 years older in one second*".

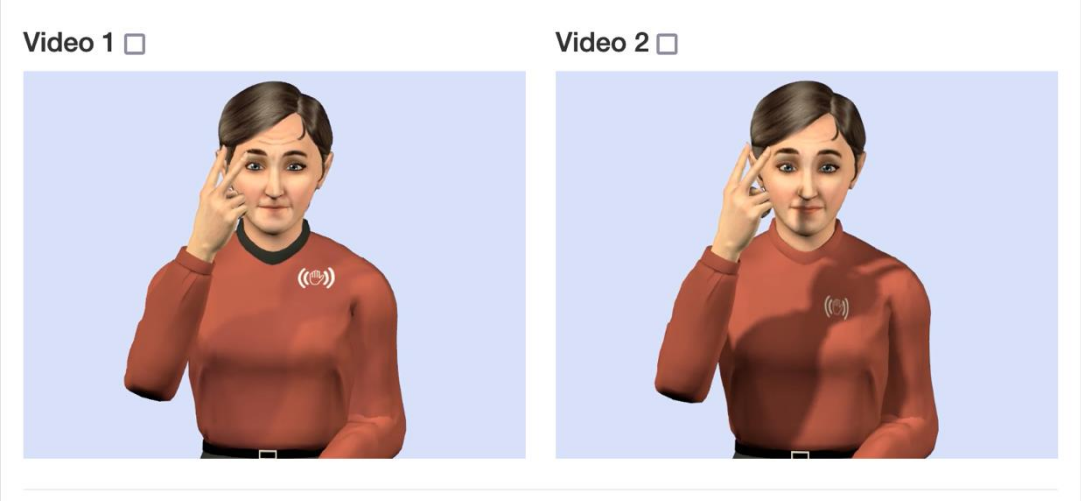


FIGURE 4.8: DGS sign VERSTEHEN with overdone non-manuals

Despite improvements, most groups were not satisfied with the current state of Paula's body and facial non-manuals (including eye-gaze). Across all language groups similar concerns were raised with respect to the face: Paula needs more diverse and more intense facial expressions if she can be comparable to a human signer. Facial non-manuals also need to be linked together; for example if Paula smiles there should also be squinting of the eyes and movement of the eyebrows to create a whole-face smile. As one French hearing participant put it: "*I think the hands are well designed, but the expressions, the eyes, the eyebrows, the corners of the mouth - there's something missing.*"

Funded by the Horizon 2020
Framework Programme of the European Union

Furthermore, the facial expressions need to be tightly linked (in timing and meaning) to the manual signing, to avoid loss of information and lack of intelligibility. One example raised was coordinating eye gaze with manual signing. DSGS participants point out that Paula appears to be staring off into the distance, without moving her eye gaze to follow pointing; this incoherence between eye gaze and manual signs creates information loss and potential confusion for the viewer. Participants also recommend adding more movements of the mouth (also known as mouth gestures), such as puffing of the cheeks, or sucked cheeks to improve naturalness of signing and thus intelligibility.

With respect to body movements, participants complained that Paula is missing movements of the shoulder and torso, both during individual signs as well as between single signs. Not only does this lack of movement make Paula appear monotonous, it also contributes to loss of information, for example in discerning signs or while referencing entities. However, deaf DGS participants pointed to a positive example of body movements, during the sign ENTSCHULDIGUNG1 (forward lean), they would like to see more of this.

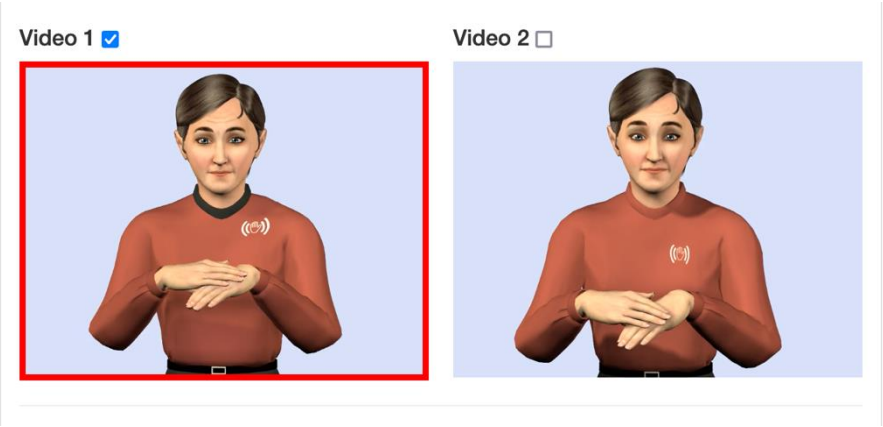


FIGURE 4.9: DGS sign ENTSCHULDIGUNG1

Participants highlighted the important role that non-manuals play a large part in the way Paula's signing conveys emotion information. On one hand, participants found Paula to lack emotional expressiveness, creating an impression of coldness; they recommend that small body movements such as forward leans or smiles can increase warmth and friendliness to the viewer. On the other hand, participants found that Paula's emotional non-manuals were at times too exaggerated to the point of inappropriateness; as one LSF participant put it, "*I really get the impression that Paula always has tears in her eyes and that she could cry any minute.*" This mismatch between non-manuals and signing also contributed to potential confusion, for example in the sentence "*Merci de patienter*", the sign PATIENTER (Figure 4.10b) is polysemous meaning both *pain* and *patience*. Paula's exaggerated facial expression suggested the meaning *pain*, when in fact the intended meaning was *patience*. Despite dissatisfaction in other groups, GSL hearing participants were pleased with Paula's affect features.

Funded by the Horizon 2020
Framework Programme of the European Union

FIGURES 4.10a and 4.10b: Excerpts from the sentence "I'm sorry I don't understand" in LSF with exaggerated NMs

4.2.5 Mouthing

Participants had several comments on the mouthing produced by Paula. While some participants appreciated that Paula's signing incorporated both mouthing and mouth gestures, the consensus was that both still needed refinement. Both DSGS and GSL deaf groups found both mouthing and mouth gestures to be not clear or not obvious enough. DGS deaf participants also pointed out to make the distribution of the two mouth actions more natural, it is necessary to follow the grammatical patterns of the language, for example relying more on mouthing in the case of polysemous signs and making use of mouth gestures in connection to verbs or intensifiers.

Another way to improve the naturalness of Paula's mouthing that was suggested was reduction; mouthing only specific syllables instead of the entire word. LSF participants found that mouthing of entire words was too much and began to resemble sign supported speech instead of natural fluid signing. Instead, deaf participants suggested that mouthing should be reduced to match normal mouthing patterns; "*in LSF we don't say all the word. I never mouth «BON-JOUR», we just mouth «B» with a smile.*"

Similarly, across both deaf and hearing group DGS participants found the mouthing to be too extreme and exaggerated, making Paula uncomfortable to look at. DGS deaf participants reported having to block out the mouthing to understand the signing. Like LSF, DGS participants expressed a desire for reduced mouthing with not all syllables, to follow more natural language use more closely. Nevertheless, DGS hearing participants highlighted one positive example; when Paula signs NOCHMAL GEBÄRDEN (*Could you please repeat that*), the mouthing of "*nochmal*" stretches across both manual signs which was considered very close to real DGS.

DSGS groups also found the mouthing, which is essential to understanding the DSGS signing, difficult to understand and affecting legibility. GSL participants also mentioned that clearer mouthing would represent an improvement for Paula's signing. Nevertheless, some participants appreciated that Paula used a lot of mouthing, indicating that this is an area in which personal preferences and potential differences in mouthing practices across European signing communities may play a large part in acceptability.

4.2.6 Methodological feedback

In this evaluation, participants were specifically asked to compare the avatar to a human signing the same content, with a video of the human signer preceding the two avatar videos. This method had pros and cons; while it ensured that participants understood the source content, it also primed participants' responses to the avatar. DGS deaf participants reported

Funded by the Horizon 2020 Framework Programme of the European Union

for example that some of the avatar's signs such as HAPPY-TIPPEN (in Sentence 5) would not have been intelligible without the human signer. However, the human video also primed participants to rate the avatar more harshly, and specifically signing that closer in style to that of the human signer was rated as better. Several groups would have preferred a rating task with no human signer, and instead for example a written prompt, as this would have helped to assess avatar quality in its own right.

With respect to rating the new and old videos side by side, while this gave a good impression of Paula's improvement, there were also some issues. For example, some DSGS participants found it very difficult to see any differences between the two avatar generations. On the other hand, DGS participants noticed early on that the new version of the avatar had a different coloured collar, and this visual cue influenced ratings (Figure 4.11). While some participants appreciated the detailed rating of the videos side by side on the same page, others did not and would have preferred them on separate pages. Finally, DGS participants reported that for the sentence "*Thank you for using our service*" it was confusing and difficult to rate, as the content signed by the two avatar versions was different.

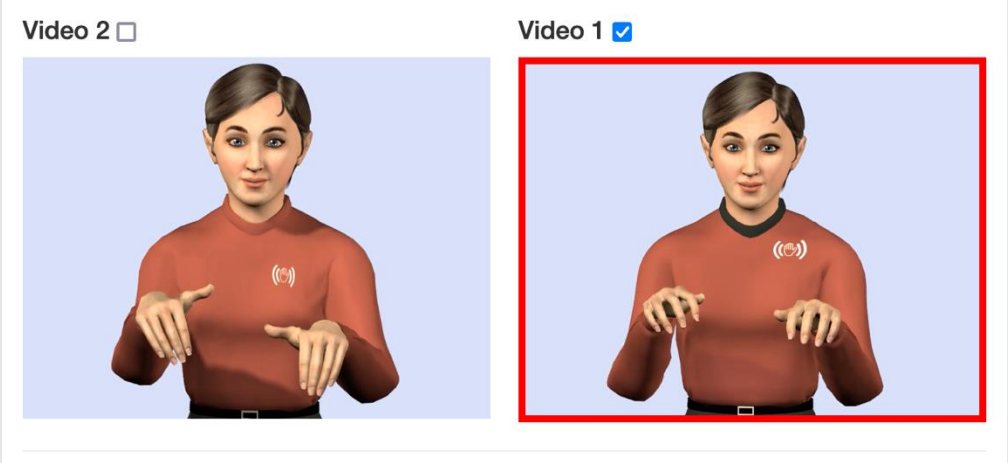


FIGURE 4.11: Differently coloured collars between old and new version of Paula

Participants reported other challenges with the method as well. In the background section of the questionnaire, DGS participants found the options to indicate where they learned sign language too limiting, as it was not possible to indicate university. The feedback form also was a source of complaint, with one DSGS deaf participant encountering issues with recording their feedback. DGS participants also found the feedback form unclear particularly the consent aspect which they thought pertained to the entire questionnaire, not the video recording. Overall, some DGS participants reported not being sure what the goal of the task was. They were unclear with how they should have rated in the questionnaire and found the open discussion too unstructured to give feedback. On the other hand, GSL participants reported high satisfaction with the task and enjoyed particularly the availability of SL instructions on each page.

From the facilitator side, one feedback point was to allow for an administrator view of the questionnaire, so that while explaining what to do they did not have to click through to select answers for all questions and could quickly skim to show participants.

Funded by the Horizon 2020
Framework Programme of the European Union

4.3 CONCLUSION

The comparison of evaluation scores between the previous and current version of the EASIER avatar technology shows clear improvement both with regards to readability and acceptance of the technology, with improved average scores across all language groups. While interpreting scores from the avatar questionnaire, it is important to note that many participants reported that seeing a human sign the same content led them to 1) rate Paula's acceptability relatively worse, as the avatar was implicitly compared to the human and 2) rate Paula's readability relatively better, as the human signer primed Paula's signs to be more intelligible. Thus, an alternative methodology may have resulted in slightly different ratings for acceptability and readability.

Qualitative feedback suggests the source of these improvements are the considerable advances with respect to fluency, mouthing and non-manuals in Paula's signing. Despite acknowledging improvements, end-users provide clear guidelines for future improvement to Paula's signing, particularly with respect to refining mouthing, enhancing non-manual features, and configuring the body position. These improvements will not only enhance acceptability, creating a warmer, more human-like avatar, but also work hand in hand to improve the readability of Paula's signing.

Funded by the Horizon 2020
Framework Programme of the European Union

5 MACHINE TRANSLATION EVALUATION

The final evaluation assessed the quality of the EASIER V2 systems on five language pairs: DGS/DE; LIS/IT; LSF/FR, BSL/EN and DSGS/DE. Each translation direction was evaluated separately. Evaluations took place fully online using the Appraise platform, a common tool for evaluating MT systems. The text to sign language translation system takes text as input and produces poses as output. The sign language to text translation system takes videos as input and produces text as output. No gloss-based systems were evaluated in this round. Sentences to be evaluated were chosen from news broadcast data, where manual alignment was available. Materials were developed by University of Zurich with support from other WP1 members for translation into local languages, and evaluations were carried out by local partners in collaboration with University of Zurich.

Funded by the Horizon 2020
Framework Programme of the European Union

5.1 METHOD

5.1.1 Participants

In line with norms in MT evaluation, we recruited trained translators and interpreters with high proficiency in the respective signed and spoken languages. For the sign to text direction, we recruited both hearing and deaf participants, and for the text to sign direction we recruited only deaf participants. Five participants were recruited per translation direction (e.g. DGS>DE had 5 participants and DE>DGS had a further 5). This resulted in a total of 41 participants, of whom 21 were deaf and 16 were hearing (4 participants did not fill out the pre-evaluation questionnaire properly so we were unable to accurately identify their background). No one participant evaluated both translation directions, however some participants took part in both the translation evaluation as well as the facilitator-led evaluation of the app and avatar.

We opted to have a relatively small number of participants complete a relatively long task for several reasons: 1) we prioritised deaf participants particularly for translation into sign language and these professionals are in short supply across the signing communities countries; 2) less participants evaluating more sentence pairs allowed us to increase statistical power of the analysis; 3) as the evaluation took place completely online it was more practical to manage fewer participants completing a longer task. Participants were recruited, managed and compensated by the local partner institutions. Table 3 summarises the participants of the MT evaluation.

Funded by the Horizon 2020
Framework Programme of the European Union

TABLE 3: Summary of participants for the online MT evaluation

Language	Partner responsible	Group	N° of participants
BSL	DCAL	BSL > EN	**5**
		EN > BSL	**4**
DGS	UHH	DGS > DE	**5**
		DE > DGS	**4**
DSGS	UZH	DSGS > DE	**5**
		DE > DSGS	**4**
LSF[13]	STXT	LSF > FR	**7**
		FR > LSF	**7**

5.1.2 Procedure

Once recruited, participants received communication directly from the responsible partner institution. The evaluation included several steps.

1. Before beginning the evaluation, they were presented with a contextualising introduction to the evaluation in text and sign language.

2. They were then invited to complete a pre-study questionnaire, which included the consent form (all information was presented in both text and sign language). The pre-study questionnaire collected demographic information about participants including their professional background, their experience with the relevant sign language, and experience evaluating MT systems.

3. Participants were then given a code to log into the evaluation platform, Appraise, and complete the core evaluation. Participants were able to work at their own pace across a timeframe of 6 weeks to complete the evaluation task. To resume after pausing work, they used their participant code and login credentials to continue their work.

[13] Two sets of LSF participants were recruited, a group from Switzerland (by STXT) and a group from France (by INT), to ensure that both signing communities were well represented.

Funded by the Horizon 2020 Framework Programme of the European Union

4. Participants then complete a post-study questionnaire, in which they were asked to give more detailed feedback on their experience using the Appraise system, as well as the most common issues they noticed with the automatic translation they evaluated.

All together the evaluation was estimated to take 6 hours per participant.

5.1.3 Appraise evaluation protocol

Machine translation output was evaluated using a Direct Assessment method, in which participants evaluated output from different origin systems. In the case of sign to text translation, the systems were 1) human translation (HUMAN) and 2) an EMSL based model (EMSL). For text to sign, the systems were 1) human translation (HUMAN) and 2) a simple lemmatization system (simple). Since the text to sign (simple) system outputs pose sequences, we applied pose estimation to the human translations (HUMAN) and also displayed them as pose sequences. This was done to de-emphasize the difference between the poses and a real person doing the signing and have evaluators focus on the quality of the message. For more information on the translation models, see D4.3 **Final translation systems**.

Evaluators were shown either the source or the reference translation and asked to rate translation quality on a scale of 1 to 100, and given some guidance to partition the scale into discrete quality levels ranging from 1 “no meaning preserved” to 100 “perfect meaning”. See Figure 5.1 for an example of the interface.

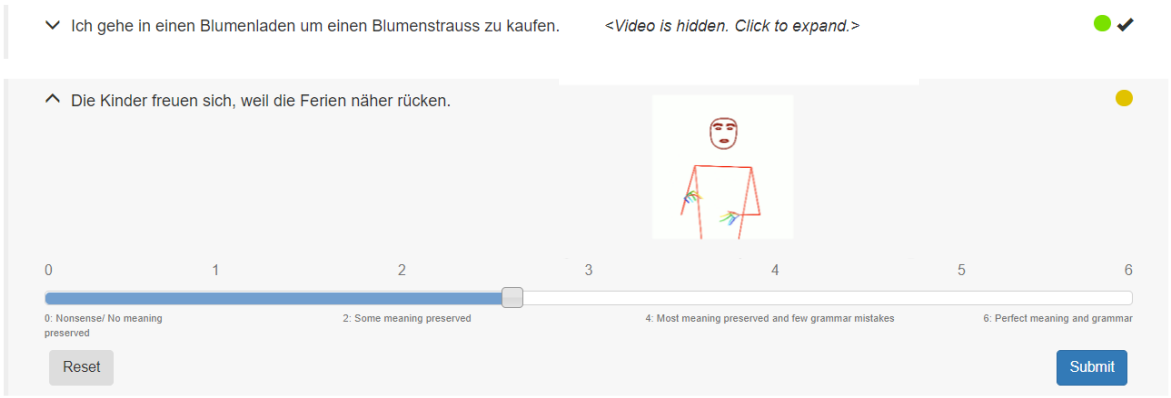


FIGURE 5.1: Example of Appraise interface for the text to sign direction

The evaluation was implemented using the tool Appraise (Federmann, 2018[14]), for which a suitable interface was created to match the needs of a sign language evaluation, supporting videos as input/output content as well as instructions. Instructions were translated into the relevant signed and spoken languages of the participating languages.

[14] Federmann, C. (2018). Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (pp. 86-88).

Funded by the Horizon 2020 Framework Programme of the European Union

5.2 EVALUATION RESULTS

The following tables show the evaluation scores for all language pairs. All systems (including the human translation) are in different quality clusters, as determined by a Mann-Whitney-U significance test.

TABLE 4: Average score given by human evaluators for all language pairs in translation direction Spoken → Signed.

Spoken → Signed		
Rank	**Average Score**	**System**
EN→ BSL		
1	42.081	HUMAN
2	20.995	simple
DE→ DGS		
1	76.077	HUMAN
2	10.341	simple
DE→ DSGS		
1	76.442	HUMAN
2	11.864	simple
FR→ LSF		
1	58.714	HUMAN
2	12.587	simple
IT→ LIS		
1	21.263	simple
2	19.717	HUMAN

Funded by the Horizon 2020 Framework Programme of the European Union

TABLE 5: Average score given by human evaluators for all language pairs in translation direction Signed → Spoken.

Signed → Spoken		
Rank	**Average Score**	**System**
BSL→ EN		
1	53.693	HUMAN
2	21.894	emsl
DGS→ DE		
1	85.842	HUMAN
2	0.286	emsl
DSGS→ DE		
1	83.096	HUMAN
2	0.753	emsl
LSF→ FR		
1	77.273	HUMAN
2	0.297	emsl
LIS→ IT		
1	25.801	HUMAN
2	7.601	emsl

Funded by the Horizon 2020
Framework Programme of the European Union

The following section outlines the main findings of the MT task:

Outlier in Italian/LIS Translation Pair: An initial observation revealed an outlier in the Italian/LIS translation pair. Human translations in this category received unexpectedly low scores. Closer analysis pinpointed the cause: misalignments in part of the Italian human-aligned news data. Consequently, we have excluded this language pair from our subsequent analyses to maintain accuracy.

Unexpectedly Low Human Translation Scores: Across all language pairs and translation directions, human translation scores were lower than anticipated. Typically, for spoken language translation, these scores hover in the high 90s percentile. However, in our findings, they ranged from 42 to 85. This discrepancy suggests a potential flaw in our methodology, particularly in the way we adapt the spoken language translation strategy. Evaluating parallel sentences in isolation, while sourcing them from contextual data, seems to be less effective for sign language.

Signed-to-Spoken Direction: The performance in the signed-to-spoken translation direction was notably poor across most language pairs. Excluding BSL→EN, the primary language pair for which EMSL was developed, our system's scores plummeted to 0 in all other language pairs. This indicates a significant challenge in the signed-to-spoken translation mechanism of our system.

Spoken-to-Signed Direction: In contrast, the spoken-to-signed translation direction showed a slightly different trend. While human translations consistently outperformed machine translations, the latter didn't score zero. Machine translations in this direction scored in the 10-20 range, suggesting a very basic but not entirely ineffective processing capability.

Funded by the Horizon 2020
Framework Programme of the European Union

5.3 POST-STUDY QUESTIONNAIRE FEEDBACK

5.3.1 Method

Many participants noted that the Appraise tool was easy to work with. However, there was a general consensus that a progress bar would have greatly improved the experience, allowing participants to better estimate the time needed for the entire task, understand how many ratings they needed to complete overall, and to have an overview of what is left to complete.

Some found the instructions clear, while others felt they needed a tutorial or a more in-depth explanation than they were offered. Some participants found particularly that the rating scale was somewhat challenging and were unsure what to choose. They would have appreciated more information such as the ratings are open to interpretation, or a few example ratings to make things clearer. In general many participants reported using the extremes of the scale, as output was either very good or very bad with little in between.

LIS participants in particular experienced issues viewing the source videos in sign to text direction; the original videos of the human interpreters often froze. And participants across all groups were affected by technical problems with loading videos throughout the evaluation. Nevertheless, participants remarked positively on the responsiveness of the EASIER partners to their questions and issues.

With respect to the pose videos, participants commented that while they were strange at first but easily got used to them. However, many found that because the hands were so colourful that it was difficult to make out hand configurations and particularly fingerspelling. One participant also lamented that the pose videos started automatically before they got the chance to read the source text - this was a waste of time as they had to wait for the (sometimes very long) video to finish before playing it again from the start.

One participant also found the setup of the task confusing with respect to how the sentences on a page related to each other. Specifically, they provided the example of the avatar using a pointing sign for referencing: the point was clearly referring anaphorically to a referent mentioned in the previous video, however it was not clear whether this should be taken into consideration or if the signing should be judged solely with respect to the sentence that it corresponded to.

5.3.2 Signed-to-Spoken

5.3.2.1 Output

Participants described the text output as either very accurate or not fitting at all. Those that were clearly machine generated presented common problems, such as repeated or rearranged segments of text, poor grammatical structure, or not being remotely related to the video; these sentences were frequently nonsensical. As one DGS participant phrased it "*the sentences make no sense. They are understandable, but they make no sense*". One participant commented that they found it a bit of a waste of time to rate so many poor translations but hoped that in the future there would be a higher quality output to evaluate.

5.3.2.2 Source content

There were a few major issues identified with the 'source' videos that affected participants ratings.

Funded by the Horizon 2020 Framework Programme of the European Union

First, participants noted that not all interpreters featured in the source videos were using fluent, grammatically correct sign language. For both LIS and LSF, participants noted that the hearing interpreters in the 'source' videos used more signed Italian and signed French, and thus it was not an accurate or representative source to be used in a task to judge translation quality. They suggested that this might in some cases artificially boost the MT quality, as the translation from e.g. signed Italian to Italian is less complicated than from LIS to Italian. They recommended that it is always better to use deaf translators/interpreters as the source of high-quality signing.

Next, participants noticed that the 'source' video was not the true original source of the information, instead the source video was a translation made from spoken language. Sometimes participants were shown the original spoken language text, and other times, the MT attempt at translating the video. This was easy to pick up on for participants because the videos came from many commonly viewed news programs. This caused problems in evaluating the sentences, giving the impression that the original spoken language text was too detailed when compared to the interpreted video. Participants describe the issue in the following quotes:

> *"With the successful items, on the other hand, I always had the feeling that they already existed BEFORE the signed translation. It didn't seem like a free translation and much more like the original text from the news. For example, the order of lists (Hungary, China, Russia...) was changed compared to the video. Why would an automatic translation make this mistake? This is much more a typical human interpreting process"*

> "*(T)he translation overshot the mark for some items and, for example, mentioned details or specified terms that could not be understood from the original text with the best will in the world (but which were certainly said in exactly the same way in the original text, the "Tagesschau" recording). I rated such over-optimized translations lower because they basically added something to the signed original text.*"

This may explain in the sign to text direction why the HUMAN (in this case the original spoken language text) receives lower scores than expected.

Finally participants reported some issues with the videos. One participant complained that the lack of context for the videos made them hard to understand and evaluate. Across groups, participants also had issues with video editing. Because the video cut off points were not always clean, participants tended to rate the translation lower. This is because even if the translation was satisfactory, it missed information from the start or end of the surrounding utterances that were contained in the video. In addition to the editing issue, LIS participants in particular experienced other technical problems with the interpreted videos such as freezing, occasional noise on video. This likely influenced LIS ratings.

5.3.3 Spoken-to-Signed

5.3.3.1 Output

Participants had a lot of feedback when it came to the pose video output (often referred to as "the avatar"). Most quickly noticed that the pose videos were either "*as reasonably close to a translation as you could expect, or nothing to do with the text*" and detected the human interpreters behind the high-quality translations. One participant remarked that it would be nice to have more differentiated examples, as they were forced to use the extremes of the scale only. Another common issue was the length of the pose videos; for those videos that contained single signs strung together, participants found it far too long to look at, with normal speed.

Funded by the Horizon 2020 Framework Programme of the European Union

Participants complained that a lot of information was missing from the pose translations, including numerals, proper nouns such as person names and locations, and fingerspelling. They also noticed that many core words were missing in the translation, suggesting a limited sign database feeding the pose videos. Participants also noted that the generated pose videos, not based on a human, did not use core aspects of signed grammar such as non-manual expressions, mouthing and role shift.

They also noticed several mistakes within the pose videos. Across all languages, participants pointed out words that the MT sometimes used the wrong signs in translation. These included several instances: 1) the MT used the sign for a word with a similar spelling but different meaning, e.g. Italian ANCÒRA versus ÀNCORA; 2) polysemous words where the MT selected the wrong meaning e.g. Italian word *sei* which translates to both the verb *to be* and the number *six* (the latter was always used); 3) this became an error carried forward creating even more confusion when some words were translated partially and incorrectly e.g. the German *Impfdose* (vaccine dose) was translated to *can* (dose); 4) the MT did not adhere to the grammatical rules of the language, e.g. not appropriately using the negative DGS morpheme when translating KANN-NICHT and instead using a standalone sign NICHT; 5) using the wrong sign that is highly similar in form to the correct translation e.g. Italian *Diciembere* translated to the sign BOVINO which is very similar in form but a totally different meaning. Other errors included pointing in inconsistent directions when referring to the same referent and switching hand dominance randomly within an utterance. One BSL participant commented that the pose videos used IS numbers instead of true BSL for numbers 1, 2 and 3 (however it was unclear if this referred to the human translator or the machine generated pose videos).

Overall comments were that participants found the pose videos totally unsuitable, using the wrong signs, missing important information and generally making no sense. Across all groups, participants were disappointed that the pose videos felt like signed speech, following word-for-word translations of the text instead of truly translating the message using the grammar of the signed languages.

While most comments had to do with the generated pose videos, participants also mentioned that for the human-based pose videos, some videos were badly cropped and cut off too quickly; this resulted in vital information not signed in the pose video and ultimately a lower score.

Funded by the Horizon 2020
Framework Programme of the European Union

6 CONCLUSIONS

The results of the MT evaluations are in line with results from previous evaluation of the EASIER translation models (D4.3) and the current state of the art of SL machine translation[15] (Müller et al, 2022[16]), where scores are consistently low. However, combining feedback from the quantitative and qualitative results of our evaluation provide valuable insights for methodological fine-tuning.

For example, while the quantitative work returns unexpectedly low human translation scores, qualitative feedback provides an explanation for this. For text-to-sign, several groups reported that the human translations created by hearing interpreters were not well formed within sign language grammar and were therefore rated poorly. For sign-to-text, participants noticed that human text translations existed before the sign language videos, and rated them poorly due to their overspecificity, including information not contained in the interpreted video.

[15] See, for example

[16] Müller, M., Ebling, S., Avramidis, E., Battisti, A., Berger, M., Bowden, R., ... & Tissi, K. (2022). Findings of the First WMT Shared Task on Sign Language Translation (wmt-slt22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 744-772).

Funded by the Horizon 2020 Framework Programme of the European Union

7 GENERAL CONCLUSIONS

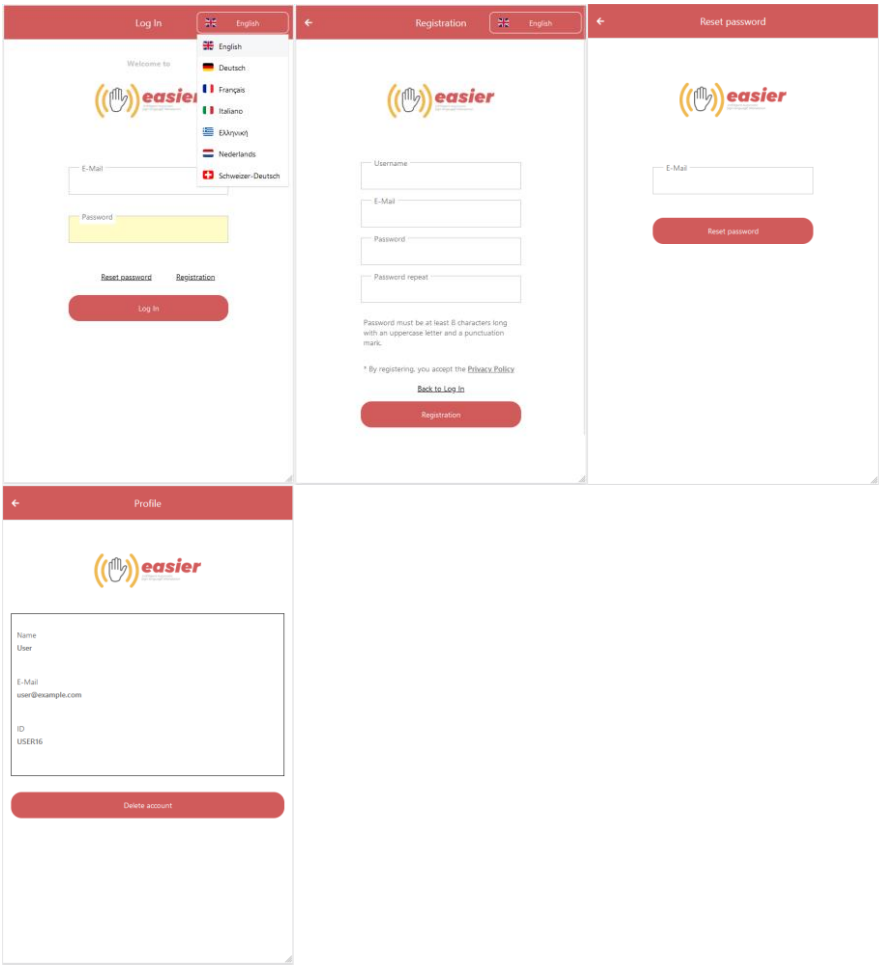
Evaluations provided valuable feedback to the EASIER project, despite coming late in the project. While all information cannot be integrated into the components in the remaining time, the results have been shared as soon as they have been collected among WP1 members (mid-end October) so that they can assess what is possible to still work on. Qualitative feedback will provide a clear path to improving these technologies in future work. Quantitative ratings also give us benchmarks to measure future improvements against.

Aside from feedback into the project, the evaluations were an important opportunity to showcase work happening inside the project to the signing communities that we hope will use these technologies. Taking place across 5 countries, including both laypeople and professionals, this was a large-scale campaign, bringing the results and the work of the EASIER project to deaf and hearing end users. For the facilitator-led groups, evaluations presented an opportunity for engagement and discussion between WP1 partners involved in the project and community members. For the MT component, conducted with deaf and hearing sign language professionals across 5 language communities, many participants expressed their excitement and enthusiasm to be included in such an evaluation campaign and interest in the new technology. This value cannot be understated as it is critical that signing communities and particularly deaf end users, feel connected and involved in the design process and are informed about the latest advancements in technology.

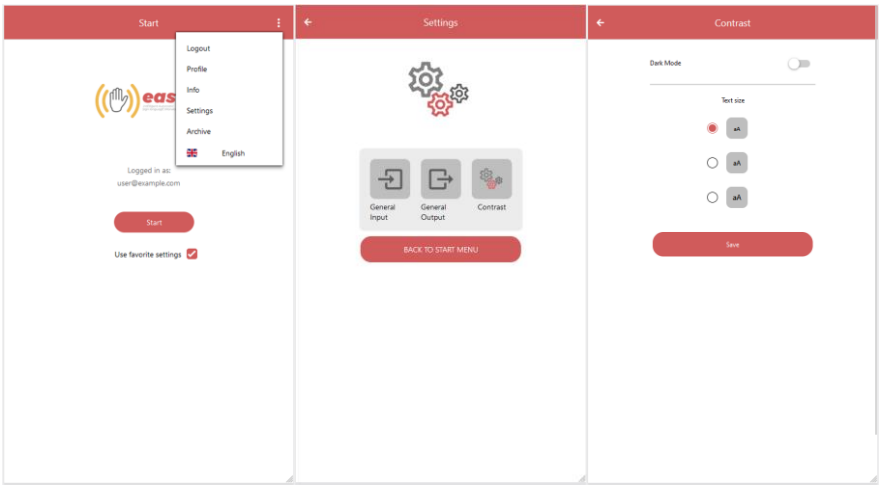
Funded by the Horizon 2020
Framework Programme of the European Union

APPENDIX A - APP PROTOTYPE

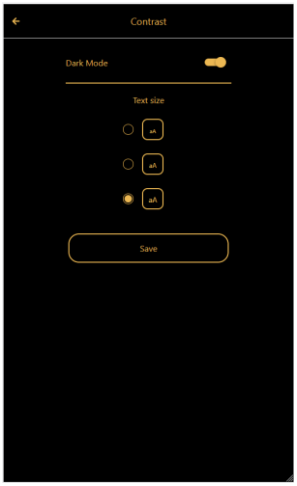
UI – User management

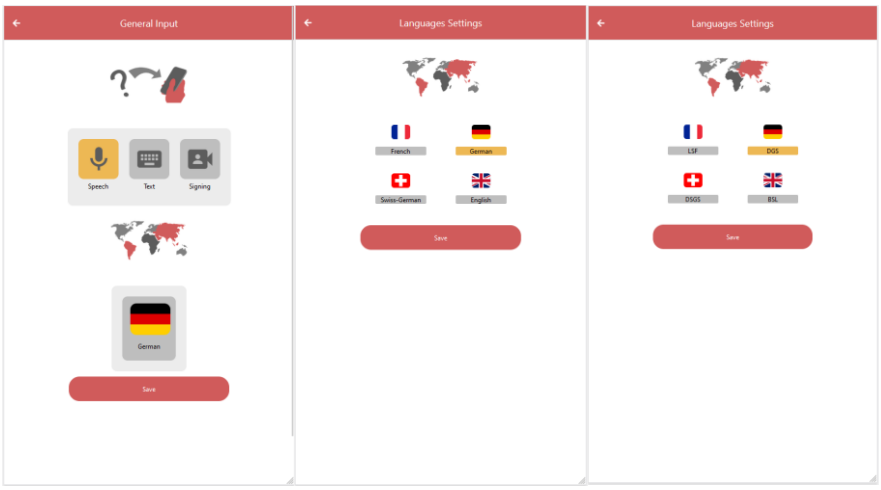


UI – Start screen and input settings



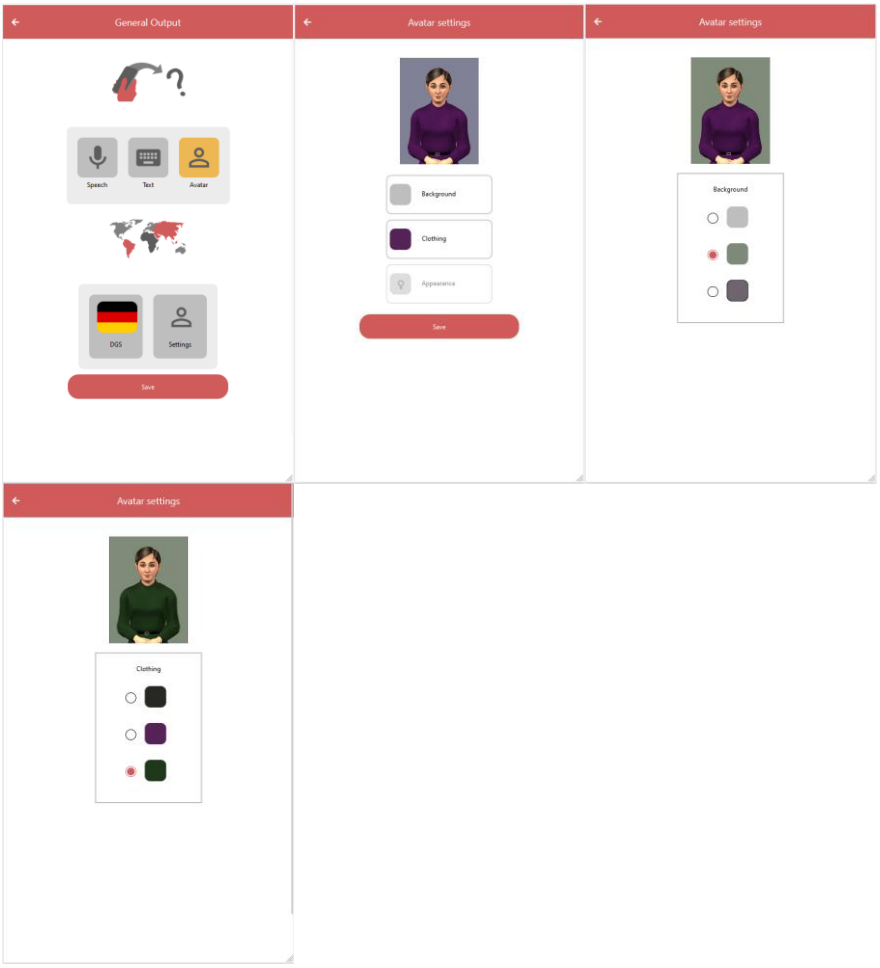
Funded by the Horizon 2020
Framework Programme of the European Union



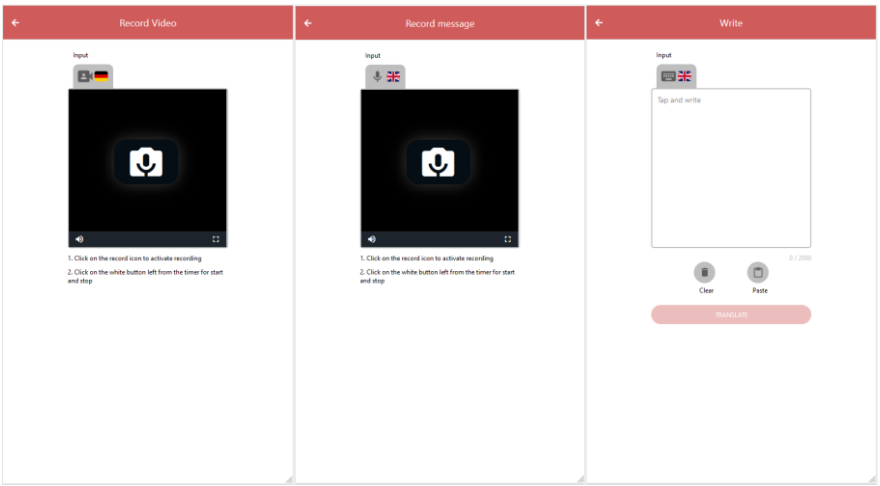


Funded by the Horizon 2020
Framework Programme of the European Union

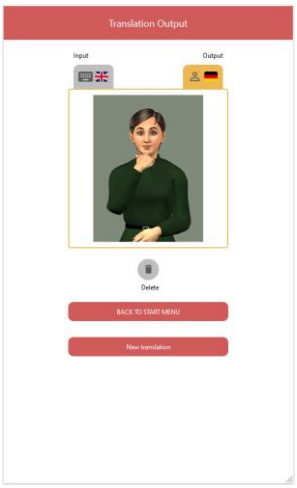
UI – Output settings



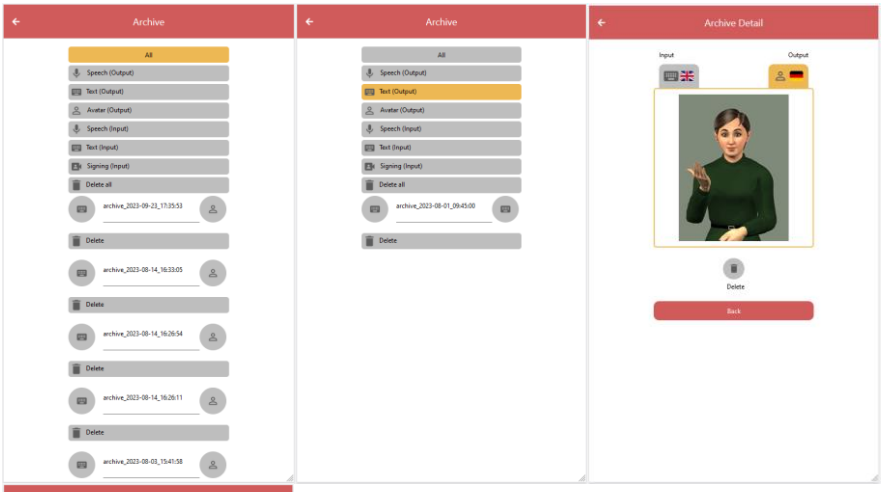
UI – Translation

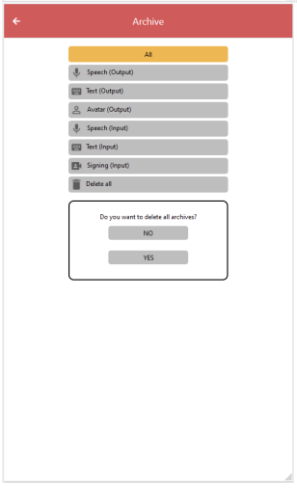


Funded by the Horizon 2020
Framework Programme of the European Union



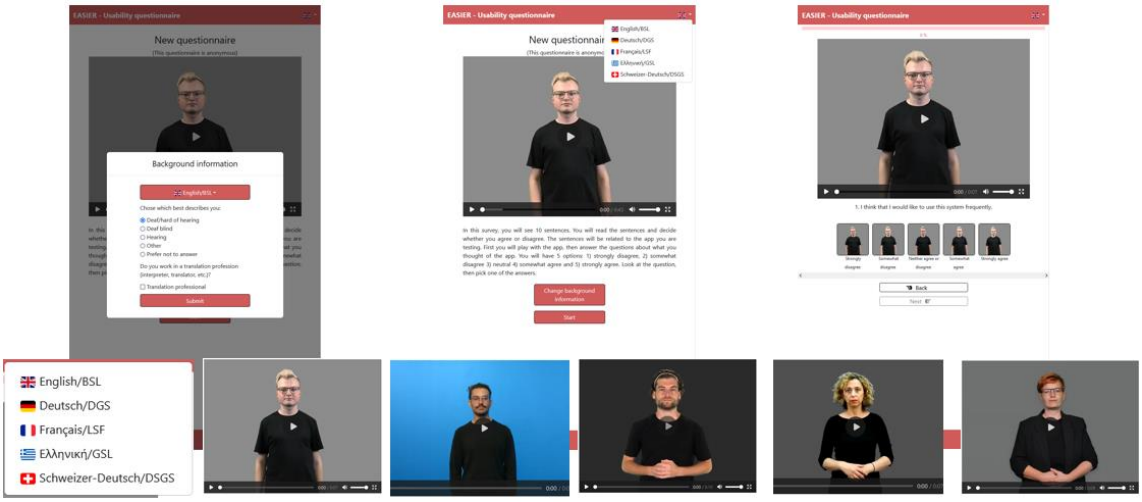
UI – Archive





Funded by the Horizon 2020
Framework Programme of the European Union

APPENDIX B – SUS QUESTIONNAIRE



Funded by the Horizon 2020
Framework Programme of the European Union