The EASIER Mobile Application and Avatar End-User Evaluation Methodology

Frankie Picron[1], Davy Van Landuyt[1], Rehana Omardeen[1],
Eleni Efthimiou[2], Rosalee Wolfe[2], Stavroula-Evita Fotinea[2],
Theodore Goulas[2], Christian Tismer[3], Maria Kopf[4], Thomas Hanke[4]

[1]European Union of the Deaf, Belgium
[2]Institute for Language and Speech Processing, ATHENA Research Center, Greece,
[3]Nuromedia GmbH, Germany
[4]Institute of German Sign Language and Communication of the Deaf, University of Hamburg, Germany
[1]{frankie.picron, davy.van.landuyt, rehana.omardeen}@eud.eu, [2]{eleni_e, rosalee.wolfe, evita, tgoulas}@athenarc.gr, [3]christian.tismer@nuromedia.com
[4]{maria.kopf, thomas.hanke}@uni-hamburg.de

Abstract

Here we report on the methodological approach adopted for the end-user evaluation studies carried out during the lifecycle of the EASIER project, focusing on the project's mobile app and avatar technologies. Evaluation was led by deaf consortium partners and performed in two cycles, involving both deaf signers and hearing sign language (SL) experts groups from five SLs to provide user feedback, which served as a reference to base the next development steps of the respective EASIER components. With this goal in mind, priorities were (i) to exploit information gathered via focus group discussions after (ii) presenting evaluators with the technological components and related questionnaires fully accessible to signers to maximize feedback and underline the importance of user involvement in the development of the technology.

Keywords: Avatar technology, end-user evaluation, sign language translation mobile application, usability, user acceptance, avatar legibility, sign language accessible questionnaire

1. The EASIER Concept

EASIER[1], a Horizon 2020 project, which ended 31st December 2023, was established with the aim to design, develop, and validate a complete multilingual machine translation system which would act as a framework for barrier-free communication among deaf and hearing individuals, as well as provide a platform to support sign language content creation.

The project concept was based on a unique combination of technological innovations, sign language resources and sign language linguistics expertise, allowing among other for the incorporation of a signing avatar that integrates sign language grammar and prosody features to perform the most advanced synthetic signing currently available, into a mobile application designed to provide users with an easy-to-use translation tool to serve everyday translation needs.[2] Envisioned functionalities of this tool included bi-directionall translation between spoken and signed languages (and vice versa), incorporating options for sign, text and speech as both input and output modalities. The EASIER mobile application was tested with five sign-spoken language pairs with the aim to create a flexible framework that could be further expanded to include other languages.

To achieve these goals, user involvement in the development of technologies has been one of the main pillars of the EASIER project. The user-centric approach of the project encompassed continuous involvement of deaf signers and SL experts in the consortium and throughout the project steps. The technology was validated in two end-user evaluation studies, the first one taking place in 2022, shortly after the mid-lifecycle of the project (see Picron, Van Landuyt and Omardeen, 2022) and the final one in 2023, close to the end of the project (see Picron et al., 2023). This paper describes in detail the design and implementation of the final end-user evaluation study of the EASIER project, specifically with respect to the signing avatar and the mobile application technologies. Our focus is documenting the evaluation methodologies in detail, rather than presenting the results, which can be found in Picron et al. (2023).

2. The EASIER End-User Evaluation Methodology

The EASIER mobile application and the avatar components were evaluated in a facilitator-led group setting, in sessions which took place both on-line and in situ, where participants were first shown the current state of the technology and asked to complete a structured rating task,

---

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

[2] An account of Machine Translation technology developed in EASIER can be found in Müller et al. (2023).

followed by a facilitator-led group discussion to get more in-depth qualitative feedback about the technology. This approach allowed us to not only get global benchmarks for how the technology is viewed by users, but also collect qualitative feedback on how to best improve technologies to achieve maximum user acceptance. For participation of end-users to all evaluation activities, a signed consent form was required, where the consent form content was provided both in text and signing.

2.1 Recruitment Strategy

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 five communities, there were two separate evaluation groups, one with deaf and hard of hearing and one with hearing participants, resulting in a total of 10 groups. Separate deaf and hearing groups were used to create a 'safe space' in which participants could freely and comfortably express themselves in their preferred language among peers. To set up the different focus groups, local project partners identified facilitators and participants for the evaluations. For the deaf groups, a deaf facilitator was chosen and for the hearing group, a hearing facilitator was chosen, while in the case of DGS the facilitator was hard-of-hearing (HoH). For 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. For each group, between 5 and 7 participants were recruited who use the target sign language. No specific professional or educational background was required for participants; however, for those evaluating the avatar, a high degree of fluency in the relevant sign language was a requirement. Recruitment was carried out through personal and professional networks, while across all groups, some participants who took part in the interim evaluation were invited back for the final evaluation. This mixture was chosen to have both, the experience of the first round allowing to judge the progress, and "fresh eyes" judging from a neutral perspective. Evaluators' anonymity was preserved since only basic demographic information was shared among technology developers, provided in the form of a cumulative report of findings from all evaluation groups.

2.2 Evaluation Setup

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 (see Fig. 1 for a group setup), some groups decided for online (see Fig. 2 for an on-line setup) or mixed evaluations to make recruitment and participation easier. Several partners conducted multiple small group evaluations to optimize 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 GoPros to record the entire scene. These recordings were then used by facilitators to later compile a report detailing the content of the focus group discussion. Recordings were kept by the local institution and not shared 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.
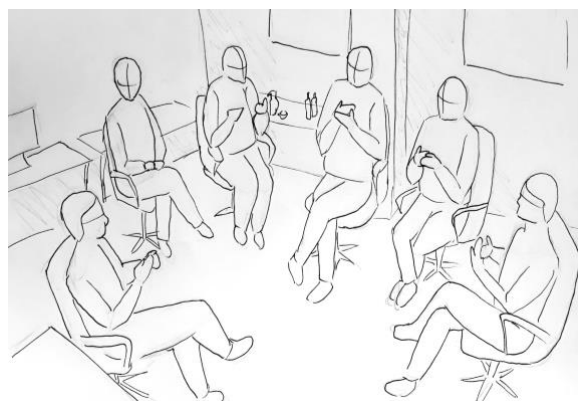

Figure 1: The EASIER group discussion setup.

Furthermore, 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.
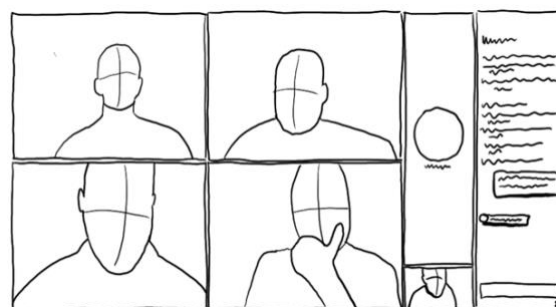

Figure 2: The EASIER on-line evaluation setup.

3. The EASIER Mobile Application and Avatar Evaluation

The evaluations consisted of two parts, one evaluating the mobile application, the other evaluating the avatar. Both evaluations consisted of an on-line questionnaire followed by in-depth discussions led by the facilitator.

3.1 The Mobile Application

The design and development of the EASIER mobile application followed a user-centric approach (Abras et al., 2004; Gulliksen et al., 2003): initial development was based on early feedback received during the user specifications and needs analysis project phase. Continuous feedback from subsequent evaluations and small working group studies with deaf users during the project lifetime guided the development of the mobile application.

The EASIER mobile app is designed to take input of either speech, sign language or text, and translate it into all of these modalities (Fig. 4).

Specific features allow users to personalize the settings for their specific input and output preferences, adjust dark and light modes, and access previous translations in an archive in the app. In the final evaluation, an early version of the mobile application was tested, which incorporated all functionalities that at a later stage supported the app's translation service. The purpose of the evaluation was therefore not to test the quality of the translation system, but instead to get feedback on the design and usability of the mobile application from the target group itself.

The final evaluation took part in three stages. Participants in the evaluation study were first instructed to create an account, and freely explore the application's features. They were then asked to complete an online questionnaire about the application's usability (see Table 1). The questionnaire was based on the System Usability Scale (SUS) (Brooke, 1995), "a reliable, low-cost usability scale that can be used for global assessments of systems usability".

For the purpose of this EASIER evaluation, following Ferreiro Lago et al. (2022), the questionnaire was presented in a bilingual format for both questions and answers (see Fig. 3), with both signed and spoken language for all five language pairs making it fully accessible to deaf evaluators. The group then came together for a discussion which concentrated on major themes regarding the application. These themes were selected based on feedback received in the interim evaluation, and involved the application's (i) settings, (ii) translation, (iii) visual design, (iv) navigation, (v) video recording and (vi) avatar output.

In both evaluation cycles, the mobile application evaluation generated a lot of engaged feedback from end users. The evaluation also added new evidence regarding the ways user preferences and expectations are formulated when participants are asked to judge the usability of a mock-up or experience the use of a prototype application.

The SUS Questionnaire
1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very awkward to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

Rating options for each question
- Strongly disagree
- Somewhat disagree
- Neither agree or disagree
- Somewhat agree
- Strongly agree

Table 1: The SUS Questionnaire and ratings used in the EASIER application evaluation.
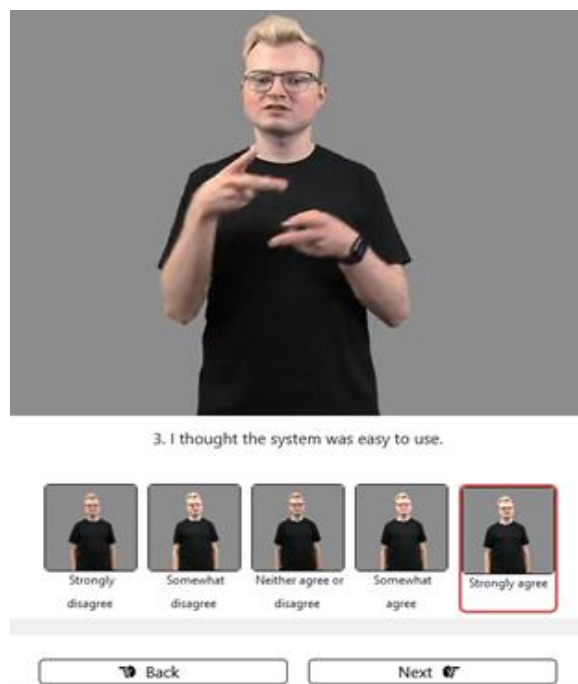

Figure 3: The SUS questionnaire for English and BSL.

While the participants appreciated the variety of input and output options for translation directions, with a mostly straightforward translation process, they demanded some advances. Main points were simplified settings, retranslation feature, searchable archive, enhancements for the video layout (e.g. mirroring, orientation), side by side in- and output (see Picron et al., 2023). These results and a median SUS score of 65 provide a useful benchmark for future work, while the qualitative round of feedback provided useful information on the strengths and weaknesses of the application, providing a roadmap for fine tuning.
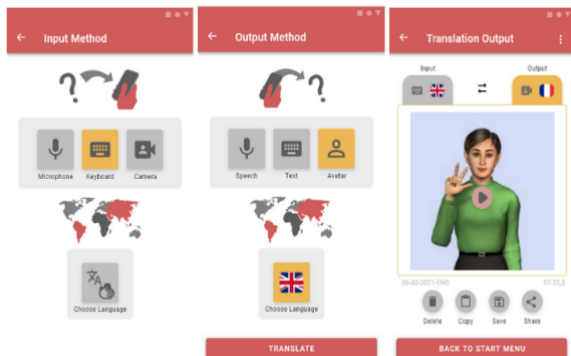

Figure 4: The EASIER app functionalities design for user input/output and translation display.

3.2   The Avatar

The main goal of avatar development for the EASIER project was to create fully legible synthetic signing (Wolfe et al., 2022a), with an avatar that was able to incorporate non-manuals, mouthing (Wolfe et al., 2022b), affect, prosody and SL grammar features beyond morphology (Hanke et al., 2023). To test the stages of development and ensure that research work was on the right track, user involvement has been critical. After basing initial work on the user needs analysis conducted in the first project phase continuous ongoing feedback was sought from the signing communities. To reach the avatar users, an on-line multilingual questionnaire[3] was developed, designed to be fully accessible via SL and easily modifiable with respect to content (Dimou et al., 2022b) (see Fig. 5). This questionnaire was used initially in a pilot survey on user preferences, drawing on two well-known avatar engines used in dynamic synthetic signing: the Anna and Paula avatars respectively (Dimou et al., 2022a). It was then adapted for use in the first and second evaluation cycles adding new content for evaluation. Although the questionnaire could be completed anonymously, it also allowed for the option of direct user input via signing into the camera of the user's device (PC/mobile phone), if this was desired.[4]
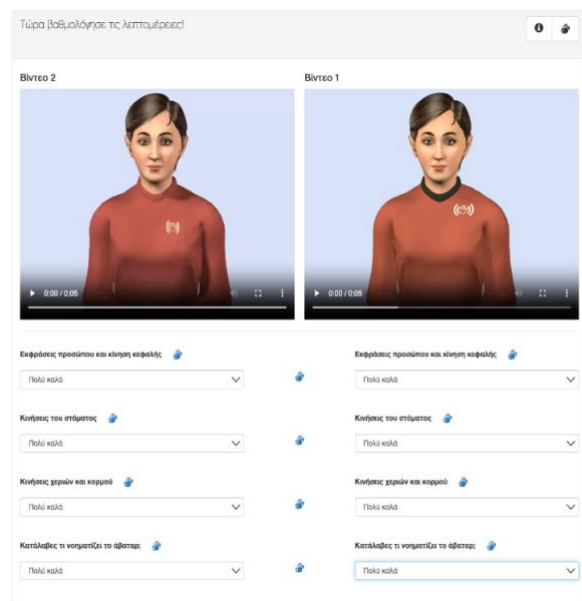

Figure 5: The EASIER avatar evaluation questionnaire (screenshot from GSL/Greek version)

The questionnaire was prepared for the four sign languages for which the avatar was available at the time: GSL, DGS, DSGS and LSF. Thus, eight groups (deaf and hearing from each of the four languages) completed the evaluation procedure. Questionnaires for each language pair were bilingual with both text and sign language and contained signed instructions for navigating each page. Before presentation of the evaluation content, some basic demographic information about the participants was collected, including their age, gender, context of sign language acquisition and self-assessment of their sign language proficiency.

For the final user evaluation, since a major goal was to measure user opinion differentiation with respect to the avatar status during the previous evaluation cycle, participants were presented with a series of screens for each animation. On the first screen, they viewed a video of an utterance produced by a human signer, which was identical in content to the utterance produced by the avatar animations. Then on the next screen, they viewed two avatar animations side by side and were asked a series of questions. Test utterances had the same semantic content across all languages.

First, they had to identify which of the two avatar animations was better. They were then asked to rate the general performance of both animations on a five-point Likert scale ranging from "very good" to "bad". On the third screen (Fig. 4) they viewed the two animations side by side again and were asked to rate each of them on (1) facial expressions and head movements, (2) mouth

---

[3] The current version of the EASIER avatar evaluation questionnaire: https://sign.ilsp.gr/slt-eval-2/

[4] Given that recording of participant video requires special permissions, consent for activation of this specific feature of the questionnaire is also mandatory.

movements, (3) hands and body, and (4) the legibility/intelligibility of the signing. All were rated on a five-point Likert scale, where the options for (1), (2) and (3) were "very good", "good", "so-so", "rather bad" and "bad", and the options for (4) were "very good", "good", "1-2 points were not clear to me", "it was difficult to understand" and "I did not understand anything" (see Figure 4, text in Greek).

The group discussions following the avatar rating focused on the overall avatar appearance and quality of the animation, the prosody in the signed utterance, the manual signing, the non-manual features and the mouthing of the animation.

In both EASIER evaluation cycles, we used the avatar evaluation questionnaire to assess the legibility and naturalness of the EASIER avatar signing. During the first evaluation cycle, although evaluators were asked to judge the avatar's hand activity only,they made clear that they wanted to see more facial activity, including mouthing as well as affect. They also wanted to see more prosodic features. These findings prioritized development during the final project period, which was evaluated at the final end-user evaluation cycle. Across all four languages evaluated, user reactions to the avatar's naturalness and legibility were positive with over 90 percent of user ratings at 3 or above (naturalness rated 3 or above: 92.3%, legibility rated 3 or above: 92.8%).

4. Conclusion

Our findings verified that continuous end-user involvement in SL technology development has proven to be the key for user acceptance and trust of the delivered tools and services. Evaluation cycles which involve larger end-user groups than those involved in a project, provide significant new feedback which is crucial to creating quantitative benchmarks to measure future improvements, while qualitative feedback provides a clear path to improving these technologies in future work. A significant aspect in evaluating SL technology is to provide evaluators with fully SL accessible questionnaires. The feedback received from the EASIER evaluator groups has verified the importance of SL based interfaces and questionnaire content.

Finally, the focus group discussion approach proved to reveal significant aspects of user attitude towards the evaluated technology, also unfolding user expectations and reservations, which the quantitative questionnaire-based approach if adopted as the only method to measure user opinion, cannot bring to light. Thus, the combination of focus group discussion and questionnaire-based evaluation can be suggested as a best practice end-user evaluation method. Finally, it must be mentioned that deaf-led evaluation is a feature that is appreciated by deaf communities.

6. References

Chadia Abras, Diane Maloney-Krichmar, Jenny Preece, (2004) User-Centered Design. In Bainbridge, W. (2004). User-Centered De-sign. In: Bainbridge, W., Ed., *Encyclopedia of Human-Computer Interaction*, Sage Publications, Thousand Oaks, CA, 445-456.

John Brooke (1995). SUS – A quick and dirty usability scale.

Athanasia-Lida Dimou, Vassilis Papavassiliou, Theodoros Goulas, Kyriaki Vasilaki, Anna Vacalopoulou, Stavroula-Evita Fotinea, and Eleni Efthimiou (2022a). What about synthetic signing? A methodology for signer involvement in the development of avatar technology with generative capacity. *Frontiers in Communication*. 7:798644.

Athanasia-Lida Dimou, Vassilis Papavassiliou, John McDonald, Theodoros Goulas, Kyriaki Vasilaki, Anna Vacalopoulou, Stavroula-Evita Fotinea, Eleni Efthimiou, Rosalee Wolfe (2022b). Signing Avatar Performance Evaluation within the EASIER Project. *7th International Workshop on Sign Language Translation and Avatar Technology (SLTAT),* 39-44.

Jan Gulliksen, Bengt Göransson, Inger Boivie, Stefan Blomkvist, Jenny Persson & Åsa Cajander (2003). Key principles for user-centred systems design. Behaviour & Information Technology, Taylor & Francis, Volume 22, 2003 - Issue 6, 397-409 | Published online: 19 May 2010.

Emilio Ferreiro Lago, María Jesús Pardo Guijarro, and Eva Gutierrez-Sigut, (2022). Diseño de cuestionarios web en investigaciones accesibles para personas sordas mediante herramientas no estándar. *Revista de Estudios de Lengua de Signos.* 4, 29-49.

Thomas Hanke, Lutz König, Reiner Konrad, Maria Kopf, Marc Schulder, Rosalee Wolfe (2023). EASIER Notation: A proposal for a gloss-based scripting language for sign language generation based on lexical data. *Eighth International Workshop on Sign Language Translation and Avatar Technology*.

Mathias Müller, Annette Rios, Amit Moryossef, Sarah Ebling (2023). EASIER final translation system v2. EASIER deliverable D4.3.

Frankie Picron, Davy Van Landuyt, and Rehana Omardeen (2022). Report on interim evaluation study. EASIER deliverable D.1.3.

Frankie Picron, Davy Van Landuyt, Rehana Omardeen, Eleni Efthimiou, Stavroula-Evita

Fotinea, Rosalee Wolfe, Theodoros Goulas, Kyriaki Vasilaki, Amit Moryossef, Mathias Müller, Sarah Ebling, and Christian Tismer, (2023). Report on final evaluation study. EASIER deliverable D1.4.

Rosalee Wolfe, John C.McDonald, Thomas Hanke, Sarah Ebling, Davy Van Landuyt, Frankie Picron, Verena Krausneker, Eleni Efthimiou, Stavroula-Evita Fotinea, and Annelies Braffort (2022a). Sign Language Avatars: A Question of Representation. *Information* 2022, 13(4), 206.

Rosalee Wolfe, John McDonald, Ronan Johnson, Ben Sturr, Syd Klinghoffer, Anthony Bonzani, Andrew Alexander, Nicole Barnekow (2022b) Supporting Mouthing in Sign Languages: New Innovations and a Proposal for Future Corpus Building. *7th International Workshop on Sign Language Translation and Avatar Technology (SLTAT),* June 2022. 125-129.