

D1.2 REPORT ON PERFORMANCE METRICS AND USER STUDY PREPARATIONS

Version: v1.0

Work Package	WP1
Task	T1.2
Due date	31/08/2022
Submission date	31/08/2022
Deliverable lead	University of Zurich
Version	1.0
Authors	Sarah Ebling, Mathias Müller
Reviewers	Rosalee Wolfe, Eleni Efthimiou (ATHENA), Christian Tismer (NURO)

Abstract	This report discusses the preparation of the user studies that constitute the EASIER v1 evaluation, the planning of the v2 evaluation, and the automatic metrics applied to continuously assess the performance of the technical components in EASIER.
Keywords	User studies, user evaluation, end users, automatic evaluation, performance metrics, subjective evaluation, objective evaluation



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 contributors
V0.1	10/08/2022	First draft	Sarah Ebling (UZH)
V0.2	22/08/2022	Submitted for review	Sarah Ebling, Mathias Müller (UZH)
V0.3	25/08/2022	Addressed review by ATHENA	Sarah Ebling, Mathias Müller (UZH)
V0.4	29/08/2022	Addressed review by NURO	Sarah Ebling, Mathias Müller (UZH)
V1.0	29/08/2022	Final version	Sarah Ebling, Mathias Müller (UZH)

DISCLAIMER

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

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

COPYRIGHT NOTICE

© 2022 EASIER Consortium

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

- * R: Document, report (excluding the periodic and final reports)
- DEM: Demonstrator, pilot, prototype, plan designs
- DEC: Websites, patents filing, press & media actions, videos, etc.
- OTHER: Software, technical diagram, etc

EXECUTIVE SUMMARY

This report discusses the preparation of the user studies that constitute the EASIER v1 evaluation, the planning of the v2 evaluation, and the automatic metrics applied to continuously assess the performance of the technical components in EASIER. The results of the v1 and v2 user evaluations will be reported in two subsequent deliverables, D1.3 and D1.4. The automatic metrics presented are used in deliverables corresponding to the individual technical components.

CONTENTS

Executive Summary	3
List of Figures	5
List of Tables	6
Abbreviations	7
1 Introduction	8
2 EASIER v1 evaluation	9
3 EASIER v2 evaluation	11
3.1 App	11
3.1.1 System Usability Scale	11
3.1.2 UI/UX assesment questions	11
3.2 Translation	12
3.2.1 Types of translation systems	12
3.2.2 Human evaluation protocols	13
3.3 Avatar comprehensibility	15
4 Automatic evaluation metrics	17
5 Conclusion and outlook	18
References	19
A Introductory texts focus group sessions	21
A.1 Overall introduction	21
A.2 Introduction app part	21
A.3 Introduction translation part	22
B Translation stimuli	23
B.1 DGS→DE	23
B.2 DE→DGS	23
B.3 BSL→EN	23
B.4 EN→BSL	23
C Questionnaire items animation	27
C.1 Question 1	27
C.2 Question 2	27
C.3 Question 3	27
D Translation: Instructions to human evaluators	30
D.1 Sign-to-spoken evaluation	30
D.2 Spoken-to-sign evaluation	30

LIST OF FIGURES

3.1	System Usability Scale for UI/UX evaluation of the EASIER app	11
3.2	Types of machine translation systems built by EASIER. pose=output of pose estimation systems. EMSL=a novel EASIER-specific representation of sign language as continuous numerical features.	13
3.3	Simplified illustration of direct assessment methods, widely used protocols for human evaluations of machine translation systems. source=input to the translation system. hypothesis=output of the system. reference=human translation.	14
3.4	Screenshot of Appraise, a browser-based tool for human evaluation of machine translation systems	15
C.1	Question 1: Does Paula sign like a human? (DSGS version)	28
C.2	Question 2: Did you understand what Paula signed? (LSF version)	28
C.3	Question 3: Did both of them [signing avatar and human signer] sign the same? (GSL version)	29

LIST OF TABLES

1.1	EASIER technical components	8
2.1	EASIER sign language/spoken language pairs	9
2.2	Focus group schedule	10
4.1	Automatic metrics for evaluation of EASIER components	17
B.1	Stimuli DGS→DE	24
B.2	Stimuli DE→DGS	25
B.3	Stimuli BSL→EN	26
B.4	Stimuli EN→BSL	26



ABBREVIATIONS

BSL	British Sign Language
DE	German
DGS	German Sign Language/ <i>Deutsche Gebärdensprache</i>
DSGS	Swiss German Sign Language/ <i>Deutschschweizerische Gebärdensprache</i>
EL	Greek
EMSL	European Meta Sign Language
EN	English
FR	French
GSL	Greek Sign Language/ <i>Ελληνική νοηματική γλώσσα (Elleniké Noematiké Glossa)</i>
IT	Italian
LIS	Italian Sign Language/ <i>Lingua Italiana dei Segni</i>
LSF	French Sign Language/ <i>Langue des Signes Française</i>
NGT	Sign Language of the Netherlands/ <i>Nederlandse Gebarentaal</i>
NL	Dutch
SQM	Scalar Quality Metric
SUS	System Usability Scale

1 INTRODUCTION

EASIER addresses the requirements of four types of users in the context of sign language/spoken language information and communication:

- Deaf EU citizens, who will be provided with a robust translation system capable of supporting their daily interaction needs
- Broadcasters, providing them with tools which accelerate robust production of high-quality SL content
- Deaf professional SL interpreters and deaf professional SL translators, boosting their efficiency in accomplishing their job tasks
- Hearing professional SL interpreters, to be supported similarly to their deaf colleagues

Within EASIER, two large user evaluation phases are planned: the first at M19-22 (v1), the second at M32/33 (v2). The aim is to evaluate the EASIER technical components listed in Table 1.1 as well as the overall EASIER system.

This report describes the user studies scheduled in EASIER as part of the v1 (Chapter 2) and v2 (Chapter 3) evaluation. Additionally, an overview of automatic metrics used to assess the performance of the individual technical components of EASIER throughout the project is given (Chapter 4).

While the deliverable at hand reports on the preparations and performance metrics used for evaluation, *conducting* the user evaluations falls under the scope of Task 1.3, with results to be reported in D1.3 (v1 evaluation) and D1.4 (v2 evaluation), respectively.

App
Signing avatar
Machine translation
Sign language (video) recognition
Affect recognition from voice
Affect recognition from text
Affect recognition from video

Table 1.1: *EASIER technical components*

2 EASIER V1 EVALUATION

Originally planned as a study with approx. 30 participants per spoken language/sign language pair in the project, the EASIER v1 evaluation has been refocused as an expert evaluation using the focus group method: Two focus group studies will be conducted for each of the spoken language/sign language pairs shown in Table 2.1.

Of the technical components shown in Table 1.1, three are assessed as part of the v1 evaluation phase: the signing avatar, the translation component, and the app. For the signing avatar, the sign languages supported by the time of the v1 evaluation are GSL, LSF, DGS, and DSGS (see Deliverable 2.1). For the translation component, the languages supported as part of the bilingual and multilingual systems delivered as D4.2 are DGS, DE, BSL, and EN. For the app, the interface will be presented to the targeted end users (see Deliverable 8.1).

For each spoken language/sign language pair, two focus groups are scheduled, one with Deaf and one with hearing participants, respectively. The Deaf participants are either sign language teachers and/or researchers or student assistants. Additionally, for the sign language/spoken language pairs that involve sign language gloss output of the machine translation system (BSL/EN, DGS/DE), participants need to be capable of reading glosses. For the hearing signers, recruitment criteria involve qualification as a sign language interpreter and/or researcher or student of sign language studies; additionally, for BSL/EN and DGS/DE, participants need to be able to read glosses. A pilot study with 1-2 participants is planned for each sign language/spoken language pair and target group (Deaf persons, hearing persons). The main study holds 4-6 participants per language pair and end user group. Facilitators are persons without direct involvement in the development of the technical component to be evaluated and ideally with knowledge of International Sign, to allow for flexibility in preparation meetings between EUD and facilitators. The coordination meetings serve the purpose to discuss, for example, anticipated questions on the side of the participants, remuneration of participants, and other practicalities.

Introductory texts for the different parts of the focus group sessions (app, translation, animation) were written in English and translated into the individual project languages to be distributed to the participants along with the informed consent forms. The English versions of the texts are shown in Appendix A.

The focus group sessions will take place physically or virtually. The animation part requires

Greek Sign Language (GSL)/Greek (EL)
French Sign Language (LSF)/French (FR)
British Sign Language (BSL)/English (EN)
Sign Language of the Netherlands (NGT)/Dutch (NL)
German Sign Language (DGS)/German (DE)
Swiss German Sign Language (DSGS)/DE
Italian Sign Language (LIS)/Italian (IT)

Table 2.1: *EASIER sign language/spoken language pairs*

Duration	BSL/EN	DGS/DE	DSGS/DE	GSL/EL	LSF/FR	LIS/IT	NGT/NL
30 mins	Welcome, introduction	Welcome, introduction	Welcome, introduction	Welcome, introduction	Welcome, introduction	Welcome, introduction	Welcome, introduction
1 hr 30 mins	Translation	Animation	Animation	Animation	Animation	App	App
1 hr 30 mins	App	Translation	App	App	App		
1 hr 30 mins		App					
Estimated time	5 hrs	7 hrs	7 hrs	5 hrs	5 hrs	2 hrs	2 hrs

Table 2.2: Focus group schedule

access to an online questionnaire and the app, to a Web interface; hence, physical focus group sessions will take place in a room with pre-installed computers or participants will be asked to bring their portable devices; Internet connection will be available.

Table 2.2 shows a sample focus group schedule for the different sign language/spoken language pairs. Following an introduction, a focus group session consists of at most three parts: a first part evaluating the signing avatar (where applicable), a second part, the translation component (where applicable), and a third part, the app.

For the app part of the v1 evaluation, the user interface and user experience design (UI/UX) based on D8.1 will be examined. Findings of the first evaluation will be integrated into D8.2. The early app version will be available at <https://easier-integration.nuromedia.com/>. Integration with other components in the first evaluation is rudimentary, a full translation will be available in the second evaluation. The users have to fulfill specific tasks to assess functionality, e.g., login, settings, and translation processes. Following this, a focus group discussion will deal with the strengths and weaknesses of the UI/UX design.

For the translation component, 18 sentences are shown for the translation direction DGS→DE, 19 for DE→DGS, 5 for BSL→EN, and 8 for EN→BSL. The stimuli are shown in Appendix B. They were chosen such that they present a basis for discussing the strengths and the weaknesses of the current version of the gloss-based translation systems.

For the animation part of the focus group sessions, the online questionnaire that will be administered to the participants is available at <http://sign.ilsp.gr/slt-eval/>. After filling in the questionnaire, participants will engage in a focus group discussion centering around the animated lexical items and sentences shown as part of the questionnaire. The items of the questionnaire are shown in Appendix C.

3 EASIER V2 EVALUATION

The second EASIER evaluation phase (v2 evaluation) to be performed in months 32/33 will assess additional as well as different aspects of the EASIER technical components. In what follows, these aspects are discussed in more detail.

3.1 APP

Focus of the v2 app evaluation is the UI/UX design including translation. The app version for the second EASIER evaluation will be a demonstrator, fully integrated with all other EASIER components and available as a web and mobile version for android and iOS delivered as D8.2. Relevant findings of the first evaluation will be included by the implementation in T8.2. Findings of this evaluation will be integrated to the final app version of D8.3.

3.1.1 System Usability Scale

Metric of the app evaluation is an adaptation of the general System Usability Scale (SUS) to this project. Each question is judged on a scale from 0 (poor, don't agree) to 7 (convincing, strongly agree) to build a SUS metric (see Figure 3.1 interpreted as a ratio scale). The questions for the app evaluation are described in Section 3.1.2.

To build a SUS for the UI/UX, several questions will be asked and the score of the answers will be averaged. Subset and classification will be taken into account for the evaluation report.

$$SUS = \frac{\sum_{i=1}^n T_i}{n}$$

3.1.2 UI/UX assesment questions

Questions described in this section will be collected by the app UI/UX questionnaire to build the SUS as described in Section 3.1.1. Questions with the letter T for “task” will be used for that average. Questions with the letter Q for “question” will be asked but excluded from the SUS, to avoid influence of the translation quality, which will be examined in a later part of the

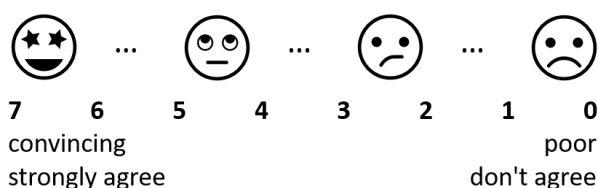


Figure 3.1: System Usability Scale for UI/UX evaluation of the EASIER app

evaluation. Further findings on that Q1 value are possible. Question with letter S is a “selection” for classification and F is “free feedback” text.

- S1: Are you a member of [] the Deaf community [] the vision-impaired community [] the community of sign language interpreters
- T1: How simple is registration to the app?
- T2: How easy was it to perform translation?
- Q1: How satisfactory is the translation quality?
- T3: How intuitive is the usage of the app?
- T4: How clear is the user interface visually?
- T5: Would you recommend the app to a friend?
- F1: Is there missing functionality?

3.2 TRANSLATION

The v2 evaluation of our translation systems will include additional system types (see Section 3.2.1) and a human evaluation of translation quality (see Section 3.2.2).

3.2.1 Types of translation systems

In the context of machine translation, sign languages can be represented in a number of ways. For instance, a sign language utterance can be represented as a video, as a sequence of transcribed glosses, as the output of a pose estimation system or as features extracted from video with machine learning methods. Different sign language representations result in different *types* of machine translation systems. Major types built in EASIER are:

- **gloss-based translation:** gloss sequences are extracted from linguistic resources such as the DGS Corpus (Hanke et al., 2020) or the BSL Corpus (Schembri et al., 2017).
- **pose-based translation:** each frame of a sign language video is represented as a *pose estimate*, a prediction of the position of various keypoints of the human body. We use two well-known pose estimation systems: Openpose (Cao et al., 2019) and Mediapipe Holistic (Lugaresi et al., 2019).
- **translation based on European Meta Sign Language (EMSL):** each frame of a sign language video is represented as numerical features extracted by an auxiliary machine learning system (see Deliverable 3.3).

See Figure 3.2 for examples. The spoken language equivalent of a sign language utterance is always represented as text.

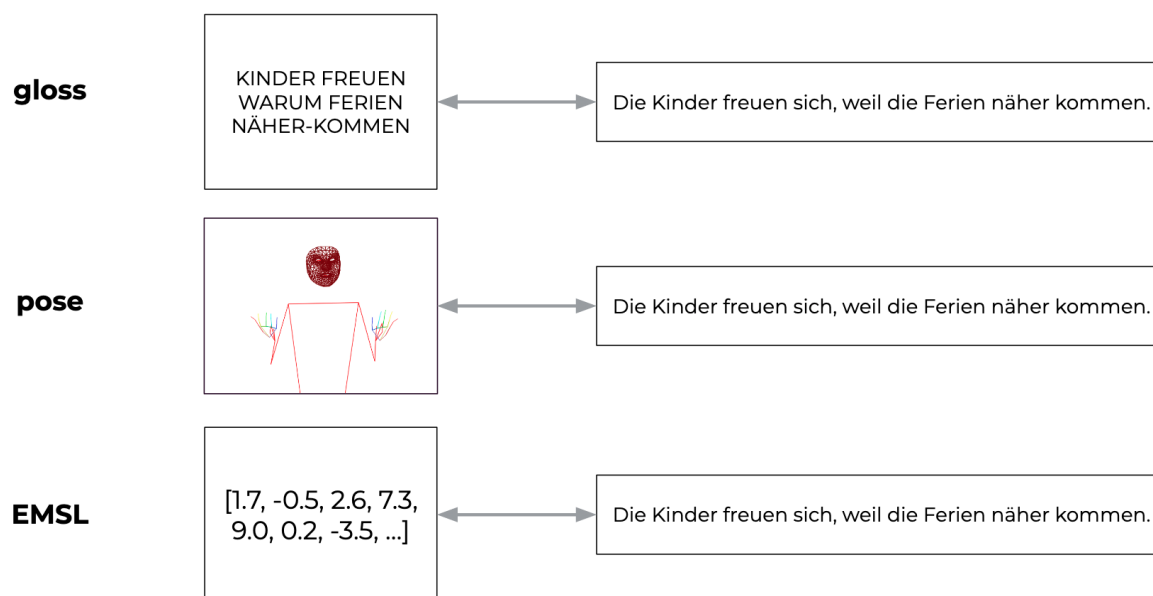


Figure 3.2: Types of machine translation systems built by EASIER. *pose*=output of pose estimation systems. *EMSL*=a novel EASIER-specific representation of sign language as continuous numerical features.

While v1 of the EASIER evaluation targeted gloss-based translation only, v2 will focus on additional types of models to be delivered as part of WP4. Specifically, models based on poses and EMSL will also be assessed.

In addition to these models involving sign languages, spoken-to-spoken translation systems will be evaluated as well.

3.2.2 Human evaluation protocols

For the v2 EASIER evaluation, the output of the translation models will be assessed by humans. Below we outline considerations for a human evaluation of sign language machine translation, as there is hardly any previous study to build upon.

Common evaluation protocols Human evaluations of machine translation output always have a *comparative* methodology, but individual methods vary in what is shown to an evaluator at any given time. The two most widely used methods are:

- **Direct assessment (DA):** One system is evaluated at any given time. The evaluator is asked to compare the MT output to either 1) the source or 2) the human reference translation. These sub-types are called *source-based* and *reference-based* DA, respectively.
- **Ranking:** several systems are evaluated at any given time. The evaluator is asked to sort system outputs by quality, producing a system ranking.

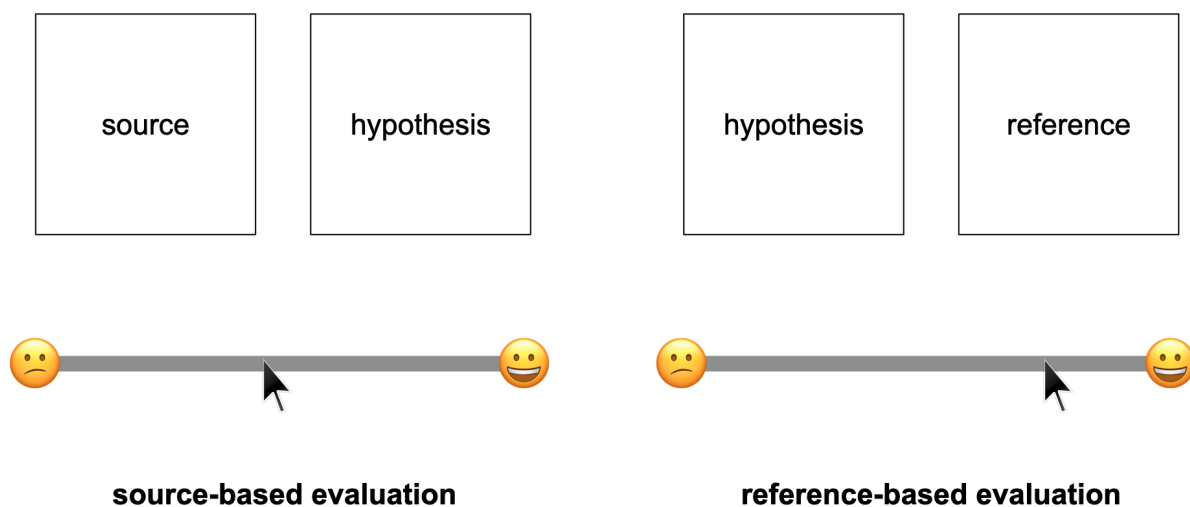


Figure 3.3: Simplified illustration of **direct assessment** methods, widely used protocols for human evaluations of machine translation systems. *source*=input to the translation system. *hypothesis*=output of the system. *reference*=human translation.

In recent years most MT evaluations have exclusively used DA methodology (Graham et al., 2016). Evaluators are shown either the source or the reference translation, and are asked to rate translation quality on a scale of 1 to 100. See Figure 3.3 for an illustration.

Design for EASIER v2 evaluation Translation quality will be assessed with source-based or reference-based DA, depending on the translation direction. We will conduct an online study using the tool *Appraise* (Federmann, 2018). As suitable user interfaces are important for such evaluations (Grundkiewicz et al., 2021), the tool was adapted to sign language in many respects.

The tool was extended to support videos as an additional modality of translation inputs or outputs and to support evaluator instructions in a sign language. See Figure 3.4 for an example of the evaluator view of *Appraise*. This new version of *Appraise* was developed for the WMT-SLT shared task on sign language translation carried out by members of EASIER¹.

Requirements for human experts The most ideal form of MT evaluation is source-based DA. Ideally, evaluators for source-based DA are bilingual, and most proficient in the target language that the MT system produces. In the context of a sign language evaluation, this means ideally that individuals are Deaf sign language users for spoken-to-sign systems and hearing sign language users with a spoken language as a first language for sign-to-spoken systems.

Deaf signers are sometimes not fully proficient in a spoken language, the spoken language being a foreign language for them. To account for this, we will run *reference-based* DA for spoken-to-sign systems, where Deaf evaluators do not need to be proficient in the corresponding spoken language.

¹<https://www.wmt-slt.com/>

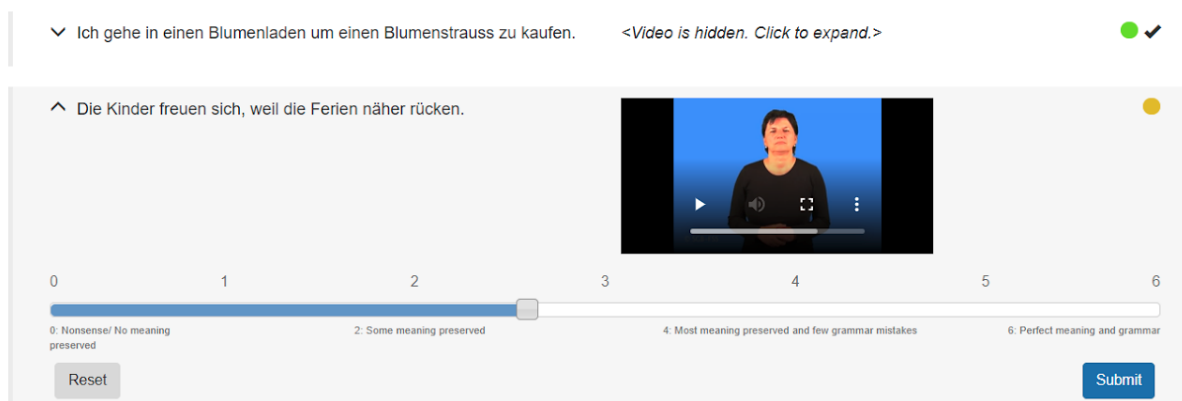


Figure 3.4: Screenshot of Appraise, a browser-based tool for human evaluation of machine translation systems

Instructions for evaluators The instructions for evaluators are adapted specifically to sign language and the new modalities (other than text) involved.

The instructions provide some guidance in the form of discrete quality levels (referred to as Scalar Quality Metric (SQM) (Freitag et al., 2021)) that partition the continuous scale of 1 to 100. The quality levels range from "0 - No Meaning Preserved" to "6 - Perfect Meaning". See Figure 3.4 for an example how the quality levels are displayed to the user.

For spoken-to-sign evaluations where the output is a sign language video (or similar), we added an evaluation criterion specific to sign languages: *naturalness of motion*. We aim to distinguish between robotic and human-like, natural motion in system outputs.

Also, following the recent evaluations at the workshop for spoken language machine translation (IWSLT 2022; Anastasopoulos et al., 2022), we remove any mention of "grammar" from the descriptions of quality levels. This was done to shift attention away from grammatical issues in the target language towards translation-breaking differences in meaning. And similar to the domain of speech, our evaluation material features continuous signing, rather than formalized signing equivalent to a written text.

The full instructions for spoken-to-sign and sign-to-spoken evaluations are included in Appendix D. We will translate these instructions to other spoken and signed languages, since Appraise also supports video instructions. We already prepared a German version of the instructions.

3.3 AVATAR COMPREHENSIBILITY

The second evaluation of the avatar will follow a format similar to the one used for the initial evaluation, namely an online questionnaire. In the questionnaire, participants will view animations created with the avatar "Paula" and evaluate them. The questions will be similar to the ones in the initial questionnaire (see Appendix C) including

1. After viewing an animation, asking the participant, "How well does the avatar sign?"
2. After viewing an animation, asking the participant, "Did you understand the signing?". The

participant has a 5-point Likert scale to rate the animation from 1 (Very well understood) to 5 (I did not understand anything).

3. In the second part of the questionnaire, a participant will see the avatar and a human side-by-side. Both the avatar and the human will sign a sentence. After viewing the animation and the video recording, a participant will answer the question, “Did both of them [avatar and human] sign the same?” with a response of “Yes” or “No”.

The second evaluation will differ in that it will introduce two additional avatars – one male, and one nonbinary. In the second evaluation every effort will be made to provide questionnaires in all seven signed languages of the EASIER project (BSL, DGS, DSGS, GSL, LIS, LSF and NGT), but this will depend heavily on identifying qualified informants in LIS, BSL, and NGT that can collaborate in creating the test questionnaires.

4 AUTOMATIC EVALUATION METRICS

Apart from the user studies described in Chapters 2 and 3, the technical components of EASIER shown in Table 1.1 are also evaluated continuously through smaller-scale user studies as well as through automatic metrics. Table 4.1 provides an overview of these metrics.

Component	Month(s)	Metric	Description
App	M23, M30	Response time, latency (Schad et al., 2010)	Tracking of response time of each EASIER component and of overall translation time
Machine translation	M16, M29	CHRf (Popović, 2015) (primary metric), BLEU (Papineni et al., 2002) (secondary metric)	Comparison of translation output against references
Sign language (video) recognition	M18, M38	WER (Morris et al., 2004) (T3.1) and BLEU (Papineni et al., 2002) (T3.2)	Validation of performance on downstream tasks such as recognition and translation with consideration given to the ability to transfer between different datasets/sign languages
Affect recognition from voice	M10, M22	Concordance Correlation Coefficient (CCC) (Lin, 1989)	Calculation of CCC on test set (no speaker overlap)
Affect recognition from text	M10, M22	F1 score (micro, macro) (primary metric), accuracy (secondary metric)	Computation of automatic metrics related to accuracy on test data
Affect recognition from video	M10, M22	Global accuracy, average per-class sensitivity; accuracy of simulated possible affective states with standardized self-reporting questionnaires (e.g. RFQ, PANAS-X)	Automated measurement on test data split from self-annotated dataset

Table 4.1: Automatic metrics for evaluation of EASIER components

5 CONCLUSION AND OUTLOOK

This report has discussed the upcoming v1 evaluation of EASIER as well as the v2 evaluation planned for the final project year, and has given an overview of automatic metrics applied to assess the performance of the EASIER technical components throughout the project. Compared to the v1 evaluation, v2 will be more extensive both in size (number of participants) and in depth (level of detail of evaluation, number of different approaches evaluated, e.g., for machine translation).

REFERENCES

- Anastasopoulos, Antonios, Loïc Barrault, Luisa Bentivogli, Marceley Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe (2022). “Findings of the IWSLT 2022 Evaluation Campaign”. In: *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*. Dublin, Ireland (in-person and online): Association for Computational Linguistics, pp. 98–157. DOI: [10.18653/v1/2022.iwslt-1.10](https://doi.org/10.18653/v1/2022.iwslt-1.10).
- Cao, Z., G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh (2019). “OpenPose: Real-time Multi-Person 2D Pose Estimation using Part Affinity Fields”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Federmann, Christian (2018). “Appraise Evaluation Framework for Machine Translation”. In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Santa Fe, New Mexico: Association for Computational Linguistics, pp. 86–88. URL: <https://www.aclweb.org/anthology/C18-2019>.
- Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey (2021). “Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation”. In: *Transactions of the Association for Computational Linguistics* 9, pp. 1460–1474. DOI: [10.1162/tac1_a_00437](https://doi.org/10.1162/tac1_a_00437).
- Graham, Yvette, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi (2016). “Is all that Glitters in Machine Translation Quality Estimation really Gold?” In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 3124–3134. URL: <https://aclanthology.org/C16-1294>.
- Grundkiewicz, Roman, Marcin Junczys-Dowmunt, Christian Federmann, and Tom Kocmi (2021). “On User Interfaces for Large-Scale Document-Level Human Evaluation of Machine Translation Outputs”. In: *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. Online: Association for Computational Linguistics, pp. 97–106. URL: <https://aclanthology.org/2021.humeval-1.11>.
- Hanke, Thomas, Susanne König, Reiner Konrad, Gabriele Langer, Patricia Barbeito Rey-Geißler, Dolly Blanck, Stefan Goldschmidt, Ilona Hofmann, Sung-Eun Hong, Olga Jeziorski, Thimo Kleyboldt, Lutz König, Silke Matthes, Rie Nishio, Christian Rathmann, Uta Salden, Sven Wagner, and Satu Wörseck (2020). *MEINE DGS. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release*. languageresource. Version 3.0. DOI: [10.25592/dgs.meinedgs-3.0](https://doi.org/10.25592/dgs.meinedgs-3.0).
- Lin, Lawrence I-Kuei (1989). “A Concordance Correlation Coefficient to Evaluate Reproducibility”. In: *Biometrics* 45.1, pp. 255–268. ISSN: 0006341X, 15410420. URL: <http://www.jstor.org/stable/2532051> (visited on 08/10/2022).
- Lugaresi, Camillo, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei



- Hua, Manfred Georg, and Matthias Grundmann (2019). “MediaPipe: A Framework for Building Perception Pipelines”. In: *CoRR* abs/1906.08172. arXiv: [1906.08172](https://arxiv.org/abs/1906.08172).
- Morris, Andrew Cameron, Viktoria Maier, and Phil D. Green (2004). “From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition.” In: *INTERSPEECH*. ISCA. URL: <http://dblp.uni-trier.de/db/conf/interspeech/interspeech2004.html#MorrisMG04>.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. DOI: [10.3115/1073083.1073135](https://doi.org/10.3115/1073083.1073135).
- Popović, Maja (2015). “chrF: character n-gram F-score for automatic MT evaluation”. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, pp. 392–395. DOI: [10.18653/v1/W15-3049](https://doi.org/10.18653/v1/W15-3049).
- Schad, Jörg, Jens Dittrich, and Jorge-Arnulfo Quiané-Ruiz (2010). “Runtime Measurements in the Cloud: Observing, Analyzing, and Reducing Variance”. In: *Proc. VLDB Endow.* 3.1–2, pp. 460–471. ISSN: 2150-8097. DOI: [10.14778/1920841.1920902](https://doi.org/10.14778/1920841.1920902).
- Schembri, Adam, Jordan Fenlon, Ramas Rentelis, and Kearsy Cormier (2017). *British Sign Language Corpus Project: A corpus of digital video data and annotations of British Sign Language 2008-2017 (Third Edition)*. London: University College London. URL: <http://www.bslcorpusproject.org>.



A INTRODUCTORY TEXTS FOCUS GROUP SESSIONS

A.1 OVERALL INTRODUCTION

EASIER is a research and innovation project funded by the European Commission within the European Union. EASIER started in January 2021 and will end in December 2023. The aim of the project is to work toward a mobile application that offers machine translation between spoken language and sign language (in both directions). The sign languages involved in EASIER are British Sign Language, French Sign Language, German Sign Language, Greek Sign Language, Italian Sign Language, Swiss German Sign Language, and Sign Language of the Netherlands. The spoken languages are those surrounding the sign languages, with Standard German shared among German Sign Language and Swiss German Sign Language.

To support translation between spoken languages and sign languages, three main language technologies are involved in EASIER: automatic sign language recognition, where the goal is to arrive at a written form of what is being signed in a video or a signing stream; automatic sign language translation, which aims to translate a written sign language representation into a written spoken language representation or the other way around; and automatic sign language synthesis, which departs from a sign language representation and displays the result by means of a signing avatar. Optionally, speech recognition and speech synthesis can be applied to turn speech input into text and text output into speech, respectively. A mobile application user interface and backend will integrate all of these components plus additional ones.

To go from sign language to spoken language, at the very least sign language recognition and sign language translation are combined. To go from spoken language to sign language, sign language translation and sign language synthesis are combined. Within EASIER, two large evaluations will take place. The first is the one you are part of, where we are not evaluating combinations of the technologies recognition, translation, and synthesis but the technologies in isolation. Specifically, for this first EASIER evaluation, we are evaluating the mobile application, the translation system, and the signing avatar.

In addition, we are only looking at some of the sign language/spoken language pairs in the project: For the translation systems, these are British Sign Language/English and German Sign Language/German. For the avatar, the sign languages considered are: French Sign Language, German Sign Language, Greek Sign Language, and Swiss German Sign Language.

A.2 INTRODUCTION APP PART

The EASIER application will allow the user to input signed (video), spoken (audio) and written (text) messages and receive a translation as signed (avatar video), spoken (synthesized audio) and written (text) in several languages.

The objective of the evaluation is to test the usability of the application. The application for this first evaluation that you are part of will be a “click dummy”, which has all user interface elements implemented but is not connected to the translation components. The avatar is implemented by a mockup response in each supported sign language with a standardized message. The

second evaluation will be a fully integrated translation application.

The first evaluation focuses on usability of registration, application/translation settings and user input to understand if the needs of the users are fulfilled.

The second evaluation will focus on the complete usage, including the translation process.

A.3 INTRODUCTION TRANSLATION PART

With regard to data, the EASIER project has two strengths: On the one hand, the project has access to three of the four largest linguistic sign language corpora in Europe (the DGS Corpus, the BSL Corpus, and the NGT Corpus). On the other hand, the consortium involves a European broadcaster with programs in three spoken languages and three sign languages as well as close cooperations with additional broadcasters. As a result of this, several thousand hours of broadcast footage for a subset of EASIER's sign and spoken languages have been collected. This data is currently being processed such that it can be used to train data-driven machine translation systems. Until this data is ready, the translation systems in the EASIER project were based on the linguistic sign language corpora in the project, specifically, the DGS Corpus and the BSL Corpus. The DGS Corpus holds about 60,000 sentence pairs to be used for translation, the BSL Corpus, about 6,000 sentence pairs. This is in relation to approx. 200 mio. sentence pairs that are estimated to be used to train Google Translate German-English. For the broadcast data available in EASIER, we are expecting to have approx. 3 mio. sentence pairs available. However, the correspondence between a signed sequence and a spoken language sentence in this data – the so-called "sentence alignment" – will be much worse than what comes from our linguistic corpora.

B TRANSLATION STIMULI

B.1 DGS→DE

B.2 DE→DGS

B.3 BSL→EN

B.4 EN→BSL

	Input	Machine translation	Human translation
1	JUNGE6 ICH2	Ich war noch jung.	Ich bin jung.
2	NETT1 EINDRUCK1 STRAHLUNG2	Er war nett und hatte eine Strahlung.	Der macht schon einen ganz netten Eindruck.
3	\$NUM-HUNDERTER1 METER1 SPRINGEN2 STARK3	Der 100-Meter-Lauf war stark.	100-Meter-Lauf und Weitsprung, darin ist er richtig stark.
4	DASSELBE2 FORMAT1 ERLEBNIS1 ICH1 DU1 DASSELBE2	Dasselbe habe ich auch erlebt.	Ich habe das auf die gleiche Weise erlebt.
5	ICH1 TAUB-GEHÖRLOS1 WARUM10	Ich fragte mich, warum ich gehörlos bin.	Warum war ich gehörlos?
6	KEIN5 PROBLEM2 FAX1 FERTIG1	Das ist kein Problem, sondern ein Fax.	Das ist gar kein Problem, es gibt ja schließlich auch Fax.
7	\$GEST-ABWINKEN1 NÄCHSTE1 TAG1 \$INDEX1 ZEITUNG1 PRESSE1	Am nächsten Tag habe ich in der Zeitung gelesen.	Am nächsten Tag haben die Zeitungen alle davon berichtet.
8	\$INDEX1 SCHÖN1 FAHRRAD1	Das ist schön, mit dem Fahrrad zu fahren.	Ist aber schön zum Fahrrad fahren.
9	ARZT1 WOLLEN2 NUR2 GELD1 VERDIENEN1 \$GEST	Der Arzt will nur Geld verdienen.	Den Ärzten geht es nur ums Geldverdienen.
10	SICH-VERSTEHEN1 HÖREND1 WIMMELN1 TRAINING1	Wir haben uns gut unterhalten und haben viel trainiert.	Ich konnte gut mit den Hörenden zusammen trainieren.
11	ICH1 FROH1 GLÜCK2 \$INDEX1 ELTERN1 ABLEHNEN2	Ich war froh, dass meine Eltern abgelehnt haben.	Ich bin froh und habe Glück gehabt, dass sie es abgelehnt haben.
12	DU1 SCHWIMMEN1 HIN-UND-HER1 ANSTRENGEND5 SEHR7	Wenn du zum Schwimmen bist, ist das anstrengend.	Wenn du hin- und herschwimmst, das ist doch unheimlich anstrengend.
13	\$PROD \$PROD VERWIRRT-SEIN1 SCHON1 ICH2 ERSTARRT1 HILFLOS1 \$GEST	Ich war völlig verwirrt, weil ich mich so verunsichert habe.	Dort war ich dann vollkommen verwirrt, ich war unsicher.
14	BEDEUTUNG1 TAUB-GEHÖRLOS1 HÖREND1 INTEGRATION1 WAS1	Das bedeutet, dass die Gehörlosen integriert werden.	Das bedeutet, dass Gehörlose und Hörende zusammen dahin gehen, oder wie?
15	UND5 \$GEST HÖREND1 ICH1 GEFÜHL3 UNTERHALTUNG1 VERGANGENHEIT1 KAUM1 SELTEN2 BEREICH1	Ich habe das Gefühl, dass ich mich kaum unterhalten habe.	Und auch mit Hörenden unterhielt ich mich nur selten.
16	ABEND2 IMMER1 FÜHREN1 \$INDEX1 ERKLÄREN1 WAS1 MORGEN1 MACHEN1 \$INDEX1 DANN1	Am Abend haben wir uns immer erklärt, was wir morgen machen sollten.	Am Abend hat der Führer immer erklärt was wir am nächsten Tag machen würden.
17	ICH1 HIER1 SCHULE1 \$NUM-TEEN1 \$NUM-ZEHNER1 \$NUM-ORD1 SCHULE1 \$GEST-OFF	Ich war in der Schule 1989 in der Schule.	Ich bin also 1950 in Köln in die Schule gegangen, das war meine erste Schule.
18	ICH2 TRINKEN-FEIN1 ZUSAMMEN3 KAFFEE2 \$GEST-ÜBERLEGEN1 TRINKEN-FEIN1 \$GEST-ÜBERLEGEN1 BECHER2 \$GEST-OFF ICH2 SCHOCK2 ERSTARRT1 ICH2 KANN1 LAUFEN8 \$GEST-OFF	Wir haben Kaffee getrunken und ich war schockiert.	Wir saßen am Tisch mit Kuchen und Kaffee und da war ich so geschockt, dass ich nicht mehr herumlaufen konnte.

Table B.1: Stimuli DGS→DE

	Input	Machine translation	Human translation
1	Eine gute Frage.	GUT1 FRAGE1	GUT1 FRAGE1 \$INDEX1
2	Kennst du das?	KENNEN1 DU1	KENNEN1 \$INDEX1
3	Das Baby war gehörlos.	BABY1 TAUB-GEHÖRLOS1	\$INDEX1 TAUB-GEHÖRLOS1 \$GEST-OFF
4	Der Mann war sofort tot.	MANN1 SOFORT2 TOD2	MANN7 TOD2
5	Du gebärdest so für OPA?	DU1 OPA6 DU1	OPA4 DU1
6	Ich wohne da in der Nähe.	ICH1 WOHNUNG1 NAHE1 \$INDEX1	ICH1 WOHNUNG1 NAHE2
7	Kein einziger Gehörloser hat dort gearbeitet.	\$NUM-EINER1 EINZIG1 TAUB-GEHÖRLOS1 ARBEITEN1 \$INDEX1	KEIN3 TAUB-GEHÖRLOS1 BEREICH1
8	Ich muss dir etwas wichtiges sagen.	WICHTIG1 MUSS1 SAGEN1 WICHTIG1	ICH1 BESCHIED1 WICHTIG1 WAS1
9	Und dann gab es natürlich viele Streitereien.	UND-DANN2 STREITEN1 \$GEST-OFF	\$INDEX1 KONFLIKT1 \$GEST-OFF
10	Es wäre schön gewesen mit Dolmetschern.	SCHÖN1 DOLMETSCHER1 \$GEST-OFF	SCHÖN1 MIT1 DOLMETSCHER1 EINSTELLEN-ARBEIT1 \$GEST-OFF
11	Wir haben auch später viel darüber gesprochen.	SPÄTER10 SPRECHEN1	SPÄTER10 ERZÄHLEN4
12	Und auch mit Hörenden unterhielt ich mich nur selten.	HÖREND1 GEBÄRDEN1 ICH1 SELTEN2 \$GEST-OFF	UND5 \$GEST HÖREND1 ICH1 GEFÜHL3 UNTERHALTUNG1 VERGANGENHEIT1 KAUM1 SELTEN2 BEREICH1
13	Der macht schon einen ganz netten Eindruck.	\$INDEX1 GUT1 NETT1 DRUCKEN-PAPIER1 \$GEST-OFF	NETT1 EINDRUCK1 STRAHLUNG2
14	Kinder zu haben, ist für Gehörlose schon anstrengender.	KIND2 ANSTRENGEND1 FÜR1 TAUB-GEHÖRLOS1 \$GEST-OFF	KIND2 TAUB-GEHÖRLOS1 MEHR1 AKTIV3 ANSTRENGEND1
15	Am nächsten Tag haben die Zeitungen alle davon berichtet.	MORGEN1 ZEITUNG1 \$GEST-OFF	\$GEST-ABWINKEN1 NÄCHSTE1 TAG1 \$INDEX1 ZEITUNG1 PRESSE1
16	Immer wieder wurde ich gestört, das war nervig.	ICH1 STÖRUNG1 \$GEST-OFF	STÖRUNG1 \$GEST-WÜTEND1 \$GEST-OFF
17	Wir konnten den Mauerfall ja vorher nicht riechen.	\$INDEX1 MAUERFALL2 KANN1 RIECHEN2	MAUERFALL1 WIR2 RIECHEN3 NICHT1
18	Viele junge Familien leben gerne in Hamburg in der Stadt.	VIEL1 JUNG1 LEBEN1 GERN1 HAMBURG1 STADT2 \$INDEX1	VIEL1 FAMILIE1 JUNG1 FAMILIE1 GERN1 IN1 HAMBURG1 STADT2 WOHNUNG2 FAMILIE1
19	Ich bin froh, dass ich heim kann und zurück an meinen Arbeitsplatz.	ICH1 FROH1 NACH-HAUSE1 ARBEITEN2 PLATZ9	FROH1 HEIM6 ARBEITEN1 PLATZ9

Table B.2: Stimuli DE→DGS

Input	Machine translation	Human translation
1 EMBARRASSED G:DISMISS	I was embarrassed.	It was awfully embarrassing, oh well.
2 MONTH FOUR02 WEEK-FOUR02 DEPTH	It was four weeks ago.	It lasted for a month, the whole four weeks.
3 LEAVE SCHOOL PT:PRO1SG FROM-TO LONDON	I left school in London.	After I had left school I had moved to London.
4 TWO INTERPRETER DSEW(1-VERT)-AT: HUMAN DSEW(1-VERT)-AT:HUMAN ONE MAN WOMAN DSEW(1-VERT)-AT:HUMAN	It was an interpreter for two years.	There were two interpreters, one male and one female at the side.
5 PT:POSS1SG BROTHER PT:POSS1SG SISTER02 PT:PRO3PL ALL HEARING SHOULD SPEECH SOME SHOULD	My brother and my sister were hearing.	Because my brother and sister are hearing, they thought we should all speak.

Table B.3: *Stimuli BSL→EN*

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

Table B.4: *Stimuli EN→BSL*

C QUESTIONNAIRE ITEMS ANIMATION

C.1 QUESTION 1

**Does Paula sign like a human?
How well does she sign?**

Very well
Well
So-so
Rather bad
Bad

Stimuli: 4 signed lexical items

C.2 QUESTION 2

Did you understand what Paula signed?

Very well
Well
1-2 points were not very clear to me.
It was difficult to understand.
I did not understand anything.

Stimuli: 5 signed sentences:

- Hello, I'm ready to begin.
- Could you repeat that?
- Sorry, I didn't understand.
- Please wait – response is pending.
- Thank you for using our service. Bye!

C.3 QUESTION 3

Did both of them [signing avatar and human signer] sign the same?

Yes/No

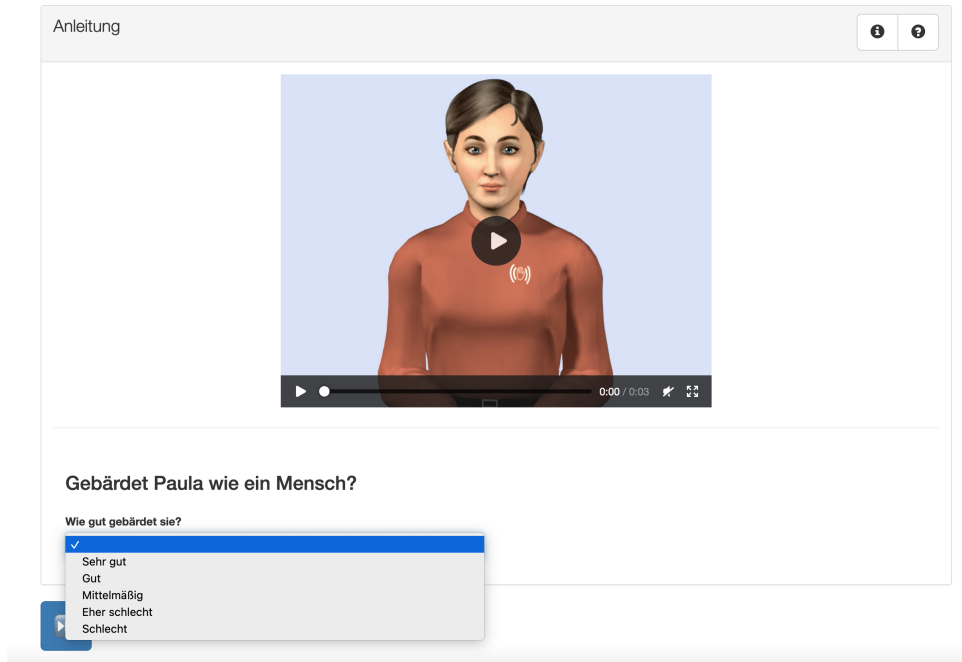


Figure C.1: Question 1: Does Paula sign like a human? (DSGS version)

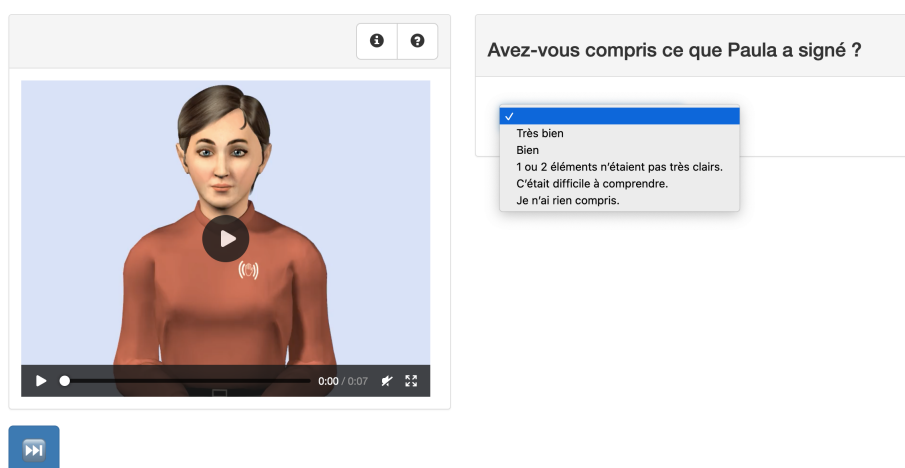
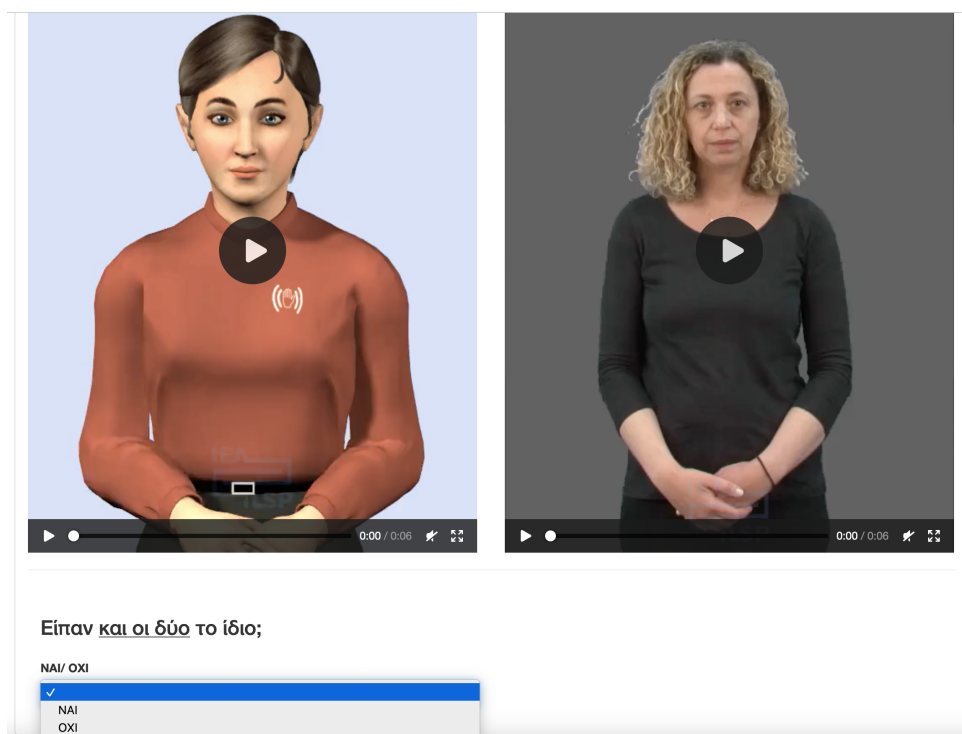


Figure C.2: Question 2: Did you understand what Paula signed? (LSF version)



Είπαν και οι δύο το ίδιο;

NAI/ OXI

✓
NAI
OXI

Figure C.3: Question 3: Did both of them [signing avatar and human signer] sign the same? (GSL version)

D TRANSLATION: INSTRUCTIONS TO HUMAN EVALUATORS

D.1 SIGN-TO-SPOKEN EVALUATION

Below you see a document with 10 sentences in Swiss-German Sign Language (Deutschschweizer Gebärdensprache (DSGS)) (left columns) and their corresponding candidate translations in German (Deutsch) (right columns). Score each candidate sentence translation in the document context. You may revisit already scored sentences and update their scores at any time by clicking at a source video.

Assess the translation quality on a continuous scale using the quality levels described as follows:

0: Nonsense/No meaning preserved: Nearly all information is lost between the translation and source. Grammar is irrelevant. 2: Some Meaning Preserved: The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Grammar may be poor. 4: Most Meaning Preserved and Few Grammar Mistakes: The translation retains most of the meaning of the source. It may have some grammar mistakes or minor contextual inconsistencies. 6: Perfect Meaning and Grammar: The meaning of the translation is completely consistent with the source and the surrounding context. The grammar is also correct.

D.2 SPOKEN-TO-SIGN EVALUATION

Below you see a document with 10 sentences in German (Deutsch) (left columns) and their corresponding candidate translations in Swiss-German Sign Language (Deutschschweizer Gebärdensprache (DSGS)) (right columns). Score each candidate sentence translation in the document context. You may revisit already scored sentences and update their scores at any time by clicking at a source text.

Assess the translation quality on a continuous scale using the quality levels described as follows:

0: Nonsense/No meaning preserved: Nearly all information is lost between the translation and source. Naturalness of motion is irrelevant. 2: Some Meaning Preserved: The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Naturalness of motion may be poor. 4: Most Meaning Preserved and Acceptable Natural Motion: The translation retains most of the meaning of the source. It may have some minor mistakes or contextual inconsistencies. Motion may appear unnatural. 6: Perfect Meaning and Naturalness: The meaning of the translation is completely consistent with the source and the surrounding context. Motion is natural.